TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

THAPATHALI CAMPUS

THESIS NO.: THA076MSISE020

AI-BASED SOCCER GAME SUMMARIZATION:
FROM VIDEO HIGHLIGHTS TO DYNAMIC TEXT SUMMARIES

BY

SUSHANT GAUTAM

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATICS AND INTELLIGENT SYSTEMS ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

KATHMANDU, NEPAL

SEPTEMBER, 2022

AI-Based Soccer Game Summarization:
From Video Highlights to Dynamic Text Summaries

by

Sushant Gautam

THA076MSISE020

Thesis Supervisor

Er. Dinesh Baniya Kshatri

A thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in Informatics and Intelligent Systems Engineering

Department of Electronics and Computer Engineering
Institute of Engineering, Thapathali Campus
Tribhuvan University
Kathmandu, Nepal

September, 2022

COPYRIGHT ©

The author has agreed that the library, Department of Electronics and Computer Engineering, Institute of Engineering, Thapathali Campus, may make this thesis work freely available for inspection. Moreover the author has agreed that the permission for extensive copying of this thesis work for scholarly purpose may be granted by the professor(s), who supervised the thesis work recorded herein or, in their absence, by the Head of the Department, wherein this thesis work was done. It is understood that the recognition will be given to the author of this thesis work and to the Department of Electronics and Computer Engineering, Institute of Engineering, Thapathali Campus in any use of the material of this thesis work. Copying of publication or other use of this thesis work for financial gain without approval of the Department of Electronics and Computer Engineering, Institute of Engineering, Thapathali Campus and author's written permission is prohibited.

Request for permission to copy or to make any use of the material in this thesis in whole or part should be addressed to:

Head
Department of Electronics and Computer Engineering
Institute of Engineering, Thapathali Campus
Thapathali, Kathmandu, Nepal

DECLARATION

I declare that the work hereby submitted for Master of Science in Infomatics and Intelligent Systems Engineering (MSIISE) at the Institute of Engineering, Thapathali Campus entitled "**AI-Based Soccer Game Summarization: From Video Highlights to Dynamic Text Summaries**" is my own work and has not been previously submitted by me at any university for any academic award. I authorize the Institute of Engineering, Thapathali Campus to lend this thesis work to other institutions or individuals for the purpose of scholarly research.

Sushant Gautam
THA076MSISE020
September, 2022

RECOMMENDATION

The undersigned certify that they have read and recommend to the Department of Electronics and Computer Engineering for acceptance, a thesis work entitled "**AI-Based Soccer Game Summarization: From Video Highlights to Dynamic Text Summaries**", submitted by **Sushant Gautam** in partial fulfillment of the requirement for the award of the degree of "**Master of Science in Informatics and Intelligent Systems Engineering**".

Thesis Supervisor
Er. Dinesh Baniya Kshatri
Assistant Professor
Department of Electronics and Computer Engineering, Thapathali Campus

---

External Examiner
Dr. Prakash Poudyal
Assistant Professor
Department of Computer Science and Engineering, Kathmandu University

---

M.Sc. Program Coordinator
Er. Dinesh Baniya Kshatri
Assistant Professor
Department of Electronics and Computer Engineering, Thapathali Campus

September, 2022

DEPARTMENTAL ACCEPTANCE

The thesis work entitled "**AI-Based Soccer Game Summarization: From Video Highlights to Dynamic Text Summaries**", submitted by **Sushant Gautam** in partial fulfillment of the requirement for the award of the degree of "**Master of Science in Informatics and Intelligent Systems Engineering**" has been accepted as a genuine record of work independently carried out by the student in the department.

Er. Kiran Chandra Dahal
Head of the Department
Department of Electronics and Computer Engineering,
Thapathali Campus,
Institute of Engineering,
Tribhuvan University,
Nepal.

September, 2022

ACKNOWLEDGMENT

This thesis work would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

First of all, I would like to express my sincere gratitude to my supervisor, **Er. Dinesh Baniya Kshatri** for providing invaluable guidance, insightful comments, meticulous suggestions, and encouragement throughout the duration of this thesis work.

I would want to convey my deepest gratitude to the Department of Holistic Systems (HOST) at Simula Metropolitan Center for Digital Engineering (SimulaMet) for giving me the opportunity to conduct this research. This work would have been impossible without the continuous support and guidance from Dr. Cise Midoglu. I would also like to thank Dr. Saeed Shafiee Sabet who helped me a lot in various stages. I am dearly obliged to Professor Pål Halvorsen for giving me an opportunity to explore the topic and providing with the resources to conduct the research.

I am also grateful to my classmates and friends for offering me advice and moral support. To my family, thank you for encouraging me in all of my pursuits and inspiring me to follow my dreams. I am especially grateful to my parents, who supported me emotionally, believed in me and wanted the best for me.

Sushant Gautam
THA076MSISE020
September, 2022

ABSTRACT

Soccer dominates the global sports market, and viewers' interest in watching videos of soccer matches is ramping up. Globally, there is a huge and constantly increasing amount of soccer game content being generated, including video footage, audio commentary, text metadata, goal and player statistics, scores, and rankings. As a large percentage of audiences prefer to follow only the major highlights of a game, the creation of multimodal (video/audio/text) summaries is of great interest to broadcasters and fans alike. In this regard, it's crucial to provide game summaries and highlights of the major game moments. However, creating summaries and annotating events most often necessitates the use of expensive equipment and a significant amount of time-consuming manual labor. Recent advancements in Artificial Intelligence (AI) technology have demonstrated great promise in this context. The purpose of this thesis is to use AI to support an automated pipeline for summarizing soccer matches. With Natural Language Processing (NLP) tools and heuristics, the emphasis is on creating comprehensive game summaries in textual form with variable length constraints, based on raw game multimedia (e.g., video and audio streams) and, where appropriate, easily accessible game meta-data. A longformer model has been fine-tuned to output a game summary for a given textual input of game captions. This work also explores the use of game audio in prioritizing game events from a summarization perspective. In particular, the Root Mean Square (RMS) audio intensity score has been extracted and used to extract the event priority to be included in the summary.

Keywords: AI, Automated Pipeline, NLP, Soccer Game Summary

TABLE OF CONTENTS

COPYRIGHT . . . . **iii**

DECLARATION . . . . **iv**

RECOMMENDATION . . . . **v**

DEPARTMENTAL ACCEPTANCE . . . . **vi**

ACKNOWLEDGMENT . . . . **vii**

ABSTRACT . . . . **viii**

TABLE OF CONTENTS . . . . **ix**

LIST OF FIGURES . . . . **xii**

LIST OF TABLES . . . . **xiv**

LIST OF ABBREVIATIONS . . . . **xv**

1 INTRODUCTION **1**

1.1 Background 1

1.2 Motivation 2

1.3 Problem Statement 3

1.4 Objectives of Thesis 4

1.5 Scope of Thesis 4

1.6 Potential Applications 5

1.7 Originality of Thesis 6

1.8 Organization of Thesis 7

2 LITERATURE REVIEW **8**

2.1 Automated Soccer Video Production 8

2.2 Video Summarization (Video Abstraction) 9

2.3 Natural Language Generation (NLG) Systems 11

2.3.1 Summarization in Journalism 11

2.3.2 Summarization in Social Media Context 13

2.4 Soccer Game Summarization 14

3 METHODOLOGY **15**

3.1 Background Concepts 15

3.1.1 Importance of Transformers in Natural Language Generation (NLG) 15
3.1.2 Basics of Transformers 16
3.1.3 LongFormer 17
3.1.4 Generative vs. Regressive Language Models 19
3.1.5 Self-Attention in Transformers 20
3.1.6 Fine-tuning GPT 22
3.1.7 Root-mean-square (RMS) Analysis 23
3.1.8 Wavelet Analysis 24
3.2 Datasets 24
3.2.1 SoccerNet Dataset 24
3.2.2 HOST Dataset 26
3.2.3 SportsSum Dataset 27
3.2.4 K-SportsSum Dataset 27
3.3 Proposed Framework 28
3.3.1 Components 28
3.3.2 Alternative Summarization Methods 33
3.3.3 Static Template Generation 35
3.3.4 Noise Suppression in Audio 36
3.3.5 Audio Intensity Analysis 37
3.4 Verification and Validation Procedures 37
3.4.1 User Study-based Evaluation 37
3.4.2 Objective Evaluation 39
4 RESULTS 42
4.1 Dataset Curation 42
4.1.1 SoccerNet 42
4.1.2 SportsSum 43
4.1.3 K-SportsSum 43
4.2 Pipeline Implementation 43
4.3 Experiments 44
4.3.1 Module 2: Noise Suppression on Extracted Audio 45
4.3.2 Module 7: Transformer Model 46
4.3.3 Module 8: Priority Filtering and Audio Dashboard 47

4.3.4 Audio Intensity Analysis 47
4.3.5 Length Control 53
5 DISCUSSION 55
5.1 Summary of Findings and Insights 55
5.2 Analysis of Results 55
5.2.1 Objective Analysis 55
5.2.2 Manual Inspection 56
5.2.3 Comparison with the State of the Art 56
5.3 Limitations 57
6 FUTURE ENHANCEMENTS 59
6.1 Open Challenges 59
6.2 Future Work 59
7 CONCLUSION 61
A APPENDIX 62
A.1 Terminology 62
A.2 Thesis Schedule 64
A.3 Literature Review of Base Papers 65
A.4 Annotation Structure of SoccerNet Metadata 69
A.5 Annotation Structure of HOST Metadata 70
A.6 Sample Summaries Generated Through the Pipeline 72
A.7 Associated Publication 76
A.7.1 Publication Acceptance Letter 77
A.7.2 Proof of Manuscript Submission 78
A.8 Plagiarism Report 79
REFERENCES 95

LIST OF FIGURES

Figure 1.1 Typical Tagging Center in Operation 1
Figure 1.2 Research Context Across Soccer Video Pipeline 5
Figure 3.1 Transformer Model Architecture 16
Figure 3.2 Masked Self-attention in Transformer Decoder Block 17
Figure 3.3 Attention Strategies for Transformers 17
Figure 3.4 Multiple Layers of Self-attention. 19
Figure 3.5 Decoder Self-Attention and Masking 21
Figure 3.6 Mother Morlet Wavelet 24
Figure 3.7 Distribution of Events in SoccerNet Dataset Meta-data 25
Figure 3.8 Per-game Distribution of Events in SoccerNet Dataset Meta-data 25
Figure 3.9 Distribution of Events in HOST Dataset Meta-data.................. 26
Figure 3.10 Per-game Distribution of Events in HOST Dataset Meta-data 27
Figure 3.11 Block Diagram of the System 28
Figure 3.12 Audio Intensity Levels for a Sample Game 30
Figure 3.13 Modules for Static Template Generation 35
Figure 3.14 The U-Net Architecture Used in Spleeter by Deezer. 36
Figure 3.15 Audio-based Filter in Summarization Pipeline 37
Figure 4.1 Audio on Temporal and Frequency Domain 45
Figure 4.2 Training Loss While Fine-tuning Longformer. 46
Figure 4.3 Overview of the Audio Dashboard 47
Figure 4.4 Audio RMS Between Events Related to Home and Away Team. 48
Figure 4.5 The Distribution of Audio RMS of Different Types of Events. 50
Figure 4.6 1-Second Audio Time Series After a Goal Event 50
Figure 4.7 Wavelet Power Spectrum 51
Figure 4.8 Global Wavelet Spectrum. 51
Figure 4.9 Power Distribution Across Wavelet Scales 52
Figure 4.10 Distribution of Length of Summary Sampled From Validation Set ... 53
Figure 4.11 Expected Vs. Generated Length of Summary 54
Figure A.1 Gantt Chart Showing the Thesis Timeline 64
Figure A.2 Exploring Attention in Trained Longformer-1 74

Figure A.3 Exploring Attention in Trained Longformer-2 74

Figure A.4 Exploring Attention in Trained Longformer-3 75

Figure A.5 LED (Longformer Encoder Decoder) Model Architecture 75

LIST OF TABLES

Table 3.1 Template for the Generation of Naive Interpretations From Metadata. . 32

Table 3.2 Alternative Methods for End-to-end Game Summarization 33

Table 4.1 Tests of Between-Subjects Effects for Audio RMS 49

Table 5.1 Objective Evaluation Using ROUGE Scores 55

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Programming Interface
ATS	Applicant Tracking System
AVI	(Audio Video Interleaved
BBC	British Broadcasting Corporation
BERT	Bidirectional Encoder Representations
BLEU	BiLingual Evaluation Understudy
CNN	Convolutional Neural Network
FFMPEG	Fast Forward Moving Picture Experts Group
FLV	Flash Video
GPT-3	Generative Pre-trained Transformer 3
GPU	Graphics processing unit
GUI	Graphical User Interface
HMM	Hidden Markov Model
JSON	JavaScript Object Notation
LOTG	Laws of the Game
LSTM	Long short-term memory
MKV	Matroska Video
ML	Machine Learning
MOV	QuickTime Movie

MP4	Moving Picture Expert Group-4
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
OTT	Over-the-Top
RMS	Root Mean Square
RNN	Recurrent Neural Network
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
STT	Speech-to-Text
SVM	Support Vector Machine
TF-IDF	Term frequency - inverse document frequency
TSN	Temporal Segment Networks
TTS	Text-to-Speech

1 INTRODUCTION

The popularity of sports broadcasting and streaming continues to rise, as does the desire to watch footage of sporting events. Nowadays, live streaming of sporting events produces the majority of video traffic and is gradually overtaking live television broadcasts [1]. For instance, according to FIFA figures, 3.572 billion people watched the 2018 FIFA World Cup [2], and as of 2020, soccer has a worldwide market share of around 45% of the $500 billion global sports business [3]. Nevertheless, the accessibility of game-related data and the sheer volume of game videos make real-time or almost real-time summaries and highlights extraction systems more crucial. Sports broadcasters are very interested in the creation of highlight reels and video recaps of games since many viewers simply want to see the key moments.

1.1 Background

Modern technologies for creating soccer highlight reels include a number of manual tasks. Figure 1.1 depicts a typical tagging centre where significant events like goals, substitutions, and cards are marked and commented on before being released as separate clips. The common pipeline consists of a "detection" step in which the video is tagged with an event such as a goal or card, relevant frames are clipped to form a highlight clip and a "refinement" phase in which the highlight clip is enhanced by tailored trimming, insertion of extra-textual descriptions and tags, and the selection/updating of a thumbnail. In addition to these operations, the production of text, audio, and video summaries from sports games is of great importance to both the public audience and broadcasters, such as news articles, social media updates, and highlight clips. However, state-of-the-art solutions in this domain are limited, and therefore subject to a lot of interest from researchers, especially in terms of automation [4]. Recent advances in artificial intelligence (AI) technology have shown enormous promise in this regard. More specifically, NLP is one of the most promising fields for the automatic generation of text summaries from videos.

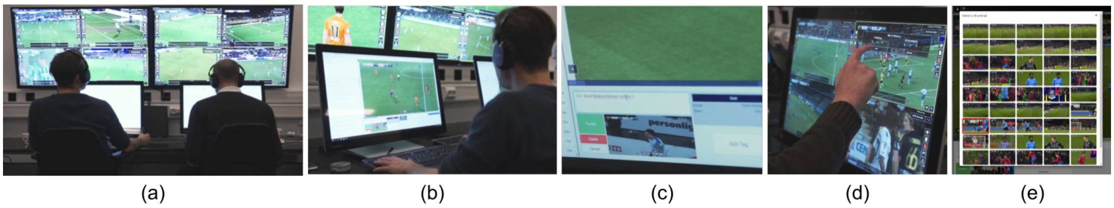


Figure 1.1: Typical Tagging Center in Operation

1.2 Motivation

The primary impetus for the creation of summarization systems is the need to handle vast quantities of data available in various forms, like videos, reports, and commentaries. Exploring lengthy information involves much time and labor. Summarization aids in combating “information overload” by emphasizing the most relevant content, minimising or removing unimportant elements. There is a lot of research around event detection in videos, event clipping for the generation of highlight clips, thumbnail selection for highlight clips, and analyzing content from social media, but not so much on the journalism or accessibility aspects of video summarization systems. Therefore, there is a dire need to establish a framework for getting text and audio summaries of videos automatically. With an emphasis on accessibility, the automatically generated audio summaries can be highly useful for visually impaired audiences as well.

Natural language understanding and generation have been interesting topics recently and there is a lot of work going on in this area. It would be intriguing to assess the capability of a successful language model that applies deep learning to create text that resembles human language, such as Longformer, for domain-specific tasks such as soccer video summarization. Combining multi-modal information of games that exist in various forms such as audio streams, metadata information about the events, and transcription of audio commentary would be a very interesting area to explore.

Building this sort of summarization system may be valuable not only for future game summaries but also for condensing past and completed games. It is a useful tool for exploring a large number of games via summaries that would have been hard to view and comprehend individually. It is exciting to conceive of a tool that may augment human skills by highlighting key facts with little investment of time, energy, and interest.

In the near future world of informatics and intelligent systems, it is natural to consider automation in practically every subject, and it is evident that sports game summarization cannot be accomplished in a manual and time-consuming manner. This definitely justifies the motivation to automate sports summarization within a concrete framework.

1.3 Problem Statement

Current state-of-art systems for soccer broadcasting and dissemination do not include automated pipelines which can summarize entire soccer games in a configurable and multimodal fashion. The process is dominated by manual efforts, and the focus of automatization has overwhelmingly been on video summaries and highlight clips.

Although it is evident from the literature that there are enough research directions on video event recognition, thumbnail selection, social media content analysis, etc., there is less on the journalistic side of sports video summarization systems. A framework for acquiring text and audio summaries from the multimodal information, which are readily available or can be feasibly generated with cutting-edge systems, is thus considered necessary. Automatically produced audio summaries would also be very beneficial for visually impaired listeners from an accessibility perspective.

Existing video summarization frameworks generate text summaries using NLP approaches. Text summaries may be generated by either processing the sampled frames in the video to identify the contents in the video or processing the sampled frames and then using NLP techniques. The subtitles/captions of sports videos, either readily available or generated from speech to text systems, are not found to have been utilized. The processing of such content from the video, if accessible, and the summarization of subtitles/captions using natural language processing algorithms provide a further possibility.

On the other hand, metadata information of games which are readily available or can be generated from event detection systems are found be be infrequently utilized in video summarization, despite the fact that they may reduce processing costs and increase continuity factor and storage efficiency.

Automatically generated text summaries have the problem of being unable to contain crucial aspects, particularly in length restricted situation. As a result, variable-length text summaries may not ideally describe the complete videos to the extent that could have been conveyed within the length restriction. In such cases, multimodal information must be exploited to as much as feasible to extract ideally relevant summaries within the text limitations.

Therefore, sports summarization system using multimodal information could be really helpful for audiences not to spend their precious time of life in watching full videos or exploring whole information about games, as well as of tremendous interest for broadcasters.

1.4 Objectives of Thesis

This thesis aims to achieve the following specific objectives:

- To design and implement an automated end-to-end pipeline for generating variable length soccer game summaries.
- To validate the pipeline using prevalent objective measures.

1.5 Scope of Thesis

Although the final system can be readily integrated into any sports video production framework and modified to generate summaries of different types of sports, this particular research focuses on the summarization of **association football (soccer)** games, and other sports are not considered.

This particular research does not focus on event detection and classification, event clipping, or highlight clip generation, but only **game summarization**. The scope of the research is clarified with Figure 1.2 with the highlighted box.

The system relies on **readily available metadata** information about the games. The assumption is that such metadata are available even in real-time during the live games from some black-box event detection systems beforehand.

The system makes no distinction between **male and female soccer** by design. No specific aspect in our pipeline has been proposed to mitigate AI bias.

The system outputs **English** language summaries. The capabilities of Longformer for summarization are explored. Text-to-Speech (TTS) systems could be then employed to convert such textual summaries to audio.

Noise suppression is needed for the videos to cancel the hootings and cheers of the crowd and make the audio of commentaries audible in adequate intensity for Speech-to-Text

(STT) systems to work. For audio extraction and processing, existing algorithms, tools, and procedures are utilized. This work does not make novel contributions in the field of noise suppression.

This research does not develop new STT models. Instead, IBM Watson is used as the STT component in the system. Longformer model pre-trained on summarization task is fine-tuned for language generation. Likewise, OpenAI Generative Pre-trained Transformer 3 (GPT-3) Application Programming Interface (API) is used for natural language understanding and metadata extraction from captions.

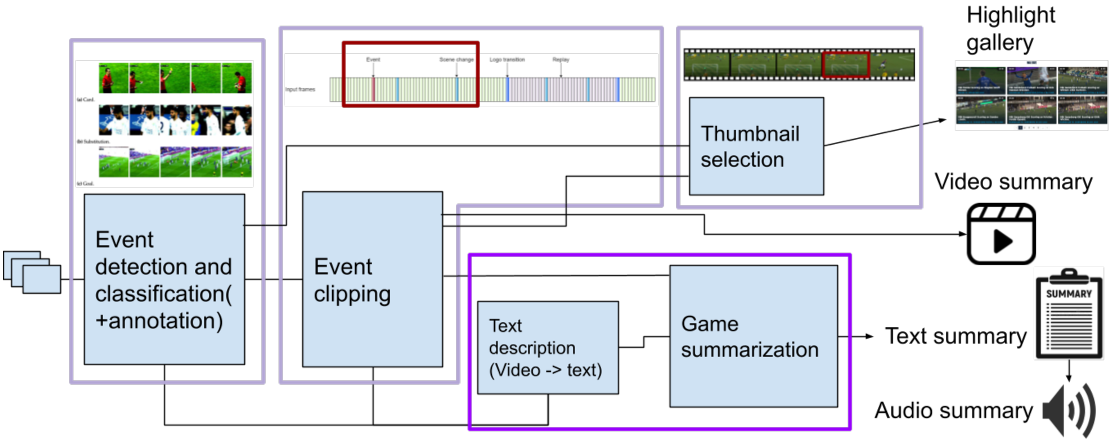


Figure 1.2: Research Context Across Soccer Video Pipeline

1.6 Potential Applications

There are a number of scenarios where the outputs of this thesis can be utilized. Below, is a list of these applications with respect to target audience.

- **Sports Companies:** This research will be directly useful to the companies like Forzasys (`https://forzasys.com/`) who have mostly manual pipelines for broadcast soccer production (example product Forzify: `https://forzasys.com/Forzify.html`). The research is also useful in analyzing subjective aspect of the summaries, for example for fans of same or opposing teams.
- **Journalism:** These kind of system can be useful to generated variable length game summaries covering important and interesting events for news reports as well as for online platforms that are fully dedicated to soccer games as well as

for soccer game section of popular news sites. The research can also be used to enhance emotional expressively in audio, which is one of the most important methods of human communication. This also explores the intrinsic aspects of human understandings and expressiveness of languages.

- **Soccer Clubs:** Systems like Reely (`https://www.reely.ai/`) and Automated Insights (`https://automatedinsights.com/`) who do automated content creations, can also benefit with automated game highlight generation and reporting, for example, to push the updated to social media in real time for sport clubs.

- **Sports Websites:** The output can be directly useful to online game portals which need to summarize the whole game in textual summary for the viewers to get the grasp of the game without watching the whole videos. Such summary can also serve as the video descriptions.

- **Web Crawlers:** Search engines are based on textual information on web content. The textual summary of the game video can be very useful in SEO and for site indexing such that a search engine can directly point the users from the natural language queries. Such efforts also relate to information distillation to address information overload.

1.7 Originality of Thesis

The research is a first hand implementation of summarizing attempt on HOST dataset. And, also probably the first attempt to incorporate audio, caption and meta-data for video summerization on SoccerNet dataset. This is the first attempt to use Text to Speech on SoccerNet and HOST dataset to automate caption generation. Such outputs can be used for future researches as well.There is a lot of work on (sports) video summarization/video abstraction, but much less on text summaries than there are on video summaries. Text summarization is only researched by itself, not together with accompanying audio summaries and meta-data. Audio summaries do not often have subjective add-ons, text summaries can be extended to text-to-speech generation with emotional effects as well.

1.8 Organization of Thesis

The thesis is structured as follows. In Chapter 1, background information is presented along with thesis objectives and applications. Furthermore, Chapter 2 contains a review of related works from the literature. In Chapter 3, the methodology to address the research question is presented, which relies on integration with multiple third-party software tools, as well as the design of original custom components. Chapter 4 contains an overview of the results and progress of the research. Chapter 5 includes an analysis of the results along with discussions. Chapter 6 contains the limitations and shortcomings of the work, as well as the various open challenges in the domain of soccer game summarization. Chapter 7 concludes the work.

The thesis schedule is included in the Appendix.

2 LITERATURE REVIEW

Researchers have been working for more than a decade to develop an automated solution for sports video semantic analysis and summarization. Various works that have been done on multiple aspects to aid such an automated pipeline are summarized below.

2.1 Automated Soccer Video Production

Event detection—also known as action detection or activity spotting—has drawn a lot of attention, recently. For instance, several two-stream Convolutional Neural Networks (CNNs) variations have been used to solve the issue [5, 6] and have now been expanded to incorporate 3D convolution and pooling [7, 8]. Temporal Segment Networks (TSN) was suggested by Wang et al. [9, 10], while C3D [11] investigated 3D convolution learning of spatio-temporal characteristics. In order to enable the network to learn spatial and temporal information independently, Tran et al. [12] adopted (2+1)D convolutions. Finding temporal points in the timeline is the goal of further techniques [13, 14, 15, 16, 17, 18, 19], however despite the fact that many of these studies show intriguing methodologies and encouraging outcomes, such technologies are not yet suitable for usage in practical deployments. The majority of the suggested models are computationally costly and only somewhat accurate, which is the cause. As long as there are no false alarms or missing occurrences, human processes are still required in deployments where action/event annotation results are utilised in an official setting (such as live sports broadcasts). The objective of these works is to accurately and efficiently automate the human event annotation procedure for soccer videos.

The amount of work that has been done in the field of *event clipping* is also limited. A shot recognition system was described by Koumaras et al. [20], and a more specialised approach was built by Zawbaa et al. [21] to handle cuts that transitioned gradually over numerous frames. Soccer video scenes were categorised by Zawbaa et al. [22] as long, medium, close-up, and audience/out of the field, and other articles showed promising scene classification findings [23, 22, 24]. Replay detection may assist in removing pointless replays from video recordings that include replays of an event. The classification names play, focus, replay, and breaks were first used by Ren et al. [25]. It has been shown that an Support Vector Machine (SVM) algorithm works well for logo-based replay detection in soccer matches, whereas an Artificial Neural Network

(ANN) is less successful [22, 21]. Additionally, it was shown that audio might be a crucial medium for locating quality cutting sites. While Tjondronegoro et al. [26] employed audio for a summarization technique, recognising whistle sounds based on the frequency and pitch of the audio, Raventos et al. [27] used audio characteristics to provide a significance score to video highlights. The topic of video segmentation in sporting events was addressed by Chen et al. [28] using an entropy-based motion technique. Lastly, other research focused on learning spatio-temporal characteristics using different Machine Learning (ML) algorithms [6, 8, 12].

These studies point to the possibility of the creation of sports videos or summaries with AI assistance. For making highlight clips, for instance, collecting temporal information might be quite helpful. The findings that are shown are still somewhat constrained, and more crucially, the real event clipping procedure is not covered. Since most of the use cases need the generation of highlight clips to be completed in real-time, computing should be possible with very low latency.

2.2 Video Summarization (Video Abstraction)

The file size of videos may be reduced in a number of ways. Hyperlapse is a method for lowering the frame rate by dropping frames [29]. Another approach is to make still images by extracting one or more keyframes. It's normal practice to pick out important sub-shots and create condensed video clips. Overall, it must meet the two requirements listed below: As much information as possible should be retained, and the number of frames or file sizes should be kept to a minimum.

Segmentation is an important process. To accelerate, Zhao et al. [30] uniformly split input films into 50-frame segments. Poleg et al. [31] propose to partition egocentric videos with complex motion using a novel Cumulative Displacement Curve and show that integrated motion vectors outperform instantaneous motion vectors. A simple technique involves calculating the colour histogram for each frame and identifying shot boundaries when the differences between two succeeding frames are below a certain threshold[32]. Potapov et al. [33] offer the KTS method for Kernel Temporal Segmentation. Pavel et al. [34] segment footage based on shot transitions and saliency maps. Amel et al. [35] recognizes shot boundaries by measuring the motion intensity between two frames across the whole movie using an adaptive rood pattern search method and then determining the threshold. Luo et al. [36] segment films depending on the motions of the camera, such as panning, zooming, and pausing. To represent temporal dependency between frames and to extract keyframes or key sub shots, Zhang et al. [37] propose two LSTM networks.

Evaluations of the value of frames and shots vary significantly. Zhou et al. [38] recommend timemapping, which is the conversion of video from high-frame-rate to low-frame-rate. In order to improve the rendering of salient motion generated by said saliency method, this paper introduces a novel saliency method, a re-timing technique to temporally resample based on frame importance generated by said saliency method, and two new temporal filters (adaptive box filter and saliency-based motion-blur filter). Sentimatic segmentation and optical flow are used in the distinctive bottom-to-top saliency technique. The results show that a saliency-based yo blur filter performs well. A pair-wise rating system is provided by Sun et al. [39] that rates the highlights of video segments without restriction and learns from online videos. First, LiveLight [30] uses a vocabulary to scan and segment the input footage. By adding each new segment that can't be partially rebuilt using the already-existing lexicon, Bengio et al. [40] builds a vocabulary of video segments. Additionally, Li et al. [41] provide a dictionary learning approach for video segmentation that takes into consideration reconstruction loss, group sparsity regularisation, patch-level and frame-level structure preservation regularizers, as well as regularizers for group sparsity.

A keyframe, several keyframes (static storyboards) [42][43], a static storyboard [44], or a shorter video clip [39] are all examples of outputs. Truong et al. [45] have completed several conventional video summaries. **Line-by-line summaries** of films are provided by novel deep learning-based video summarising techniques [46] and such techniques have been suggested since 2006. Keyframe extraction and video skimming are approached by Cong et al. [47] as a dictionary selection problem. The work by Sunet al. [48] offers a novel poselet-based saliency detector, and person recognition and tracking system. A sequential Determinantal Point Process (seqDPP) is presented by Gong et al. [49] to describe videos in a supervised manner. It also provides evaluation metrics. Given two summaries, identify the matching frames whose visual distance is below a threshold and determine their accuracy, recall, and F-score. And by greedily optimising the F-score between the ground truth and a number of human-annotated videos, it creates a summary of the ground truth for each video.

Mindek et al. [50] summarize multiplayer 3D scene games based on game rules. There are other evaluations of several types of video summaries in Sreeja et al. [51].

2.3 Natural Language Generation (NLG) Systems

Applicant Tracking System (ATS) is widely used in text mining and analytics applications like information retrieval, information extraction, question answering, etc. ATS is used in conjunction with information retrieval techniques to improve search engine performance. An algorithm and pseudo-code search engine is presented by Tuarob et al. [52]. A dataset is initially compiled by removing algorithms from academic works. The collected algorithms from scientific papers are then enhanced with more textual data using ATS. Text summarization is used by Yulianti et al. [53] to extract answers to questions that are not factoids. An extractive multi-modal summarization (MMS) technique is provided by Li et al. [54] for asynchronous collections of text, images, audio, and video. On the basis of the sources mentioned above, the suggested system generates a textual summary.

The absence of agreement among researchers around evaluation methodologies on text summary evaluation and the lack of thorough, up-to-date research makes text summarization clearly reflects the shortcomings. The popularly used n-gram-based evaluation metrics similar to ROUGE and even embedding-based metrics like BertScore are not adequate to assess the quality of summarization, according to Fabbri et al. [55]'s analysis of the current evaluation metrics used for such tasks The findings clearly reflect the need for future studies on text summarization evaluations and models.

2.3.1 Summarization in Journalism

For every game, the Dutch data-to-text system PASS [56] generates two sports reports, each with a unique tone based on the reader's team. A computer can complete this laborious task more quickly than a human journalist, giving readers a more customised and pertinent report. The input data is scraped from websites, including past game results and previous matches between the teams. The templates were made using the 2016 MeMo FC corpus [57]. It contains match reports that were specifically written for the teams that played in the game and is aimed at the supporters of those clubs. Given the anticipated personalization of the content, this makes it ideal for PASS. PASS was used to retrieve a sizable number of event categories and templates for each category. According to the authors, the algorithm produces text with a similar amount of variance as GoalGetter [58]. Corney et al. [59] advocated subjective summarization via social media fan interactions. Prior to using a subject identification algorithm to identify the sub-topic events by analysing quick increases in word frequency, they first established the user's preferred team. A commentary system that recounts the happenings in a certain match is presented by Chen et al. [60]. This already represents a significant departure from the dynamic summary generator, where only the most significant events are deemed pertinent to be included in the text and where practically every event will appear in the text. Another contrast is that the text was generated by machine learning methods rather than templates by Chen et al. [60]. The system was developed using human-commented games from the Robocup simulation league[1], and three algorithms were added to provide commentary for games that had not yet been played. In order to determine what kinds of occurrences (such as passes or goals) are most likely to be reported on by human commentators, it adopts a probabilistic technique for the content-selection problem.

The 2017 Finnish municipal elections were covered in news articles written in English, Swedish, and Finnish using an NLG template-based approach [61]. To prevent referring to the same entity by the same name more than once, it employed basic templates like "$entity won $value new seats in $location" as well as various strategies like employing a referring expression generator. Using keywords or natural language, the system created by Plachouras et al. [62] could search for financial data. The system assesses the supplied query, gathers the relevant information, and provides an NLG-based answer. The system locates the record holding India's GDP in 2010 and sends a text answer in response to the query "India's GDP 2010". It has a module that uses one of the several templates provided for each use case to construct different parts of the answer. With the use of numerical weather prediction data, SumTime-Mousam [63] generates weather forecast predictions. Since the output will be in a meteorological sublanguage rather than ordinary English, new grammatical rules were created as opposed to using templates. The majority of the essay focuses on how individuals and the system choose terms like "reversing" or "becoming" to describe a change in wind direction.

[1] https://robocup.org

The Los Angeles Times started a blog in 2007 to report on killings, and it used automated writing generated from a basic template [64]. In 2014, the same journal published the first earthquake-related article using Quakebot [65]. Using NLG technology, the BBC was able to upload news articles for each of the United Kingdom's 650 seats on the evening of the 2019 general election [66]. Automated Insights[2] and Narrative Science[3], which created WordSmith and Quill, respectively, are the most important enterprises from a monetary standpoint.

2.3.2 Summarization in Social Media Context

Several studies on social media text summarization have been conducted. Using a graph-based technique, Sharifi et al. [67] summarised the most frequently occuring phrases in a sample of tweets. On the basis of the words that had been found, a phrase was then selected to sum up the situation. Inouye et al. [68] summarised tweets using a clustering approach. The similarity metric was used to group tweets into K-clusters, and the highest-scoring tweets from each cluster were extracted and weighted using the hybrid Term frequency - inverse document frequency (TF-IDF) weighting approach. The Hidden Markov Model (HMM) was modified by Chakrabarti et al. [69] to provide a method for identifying sub-events. Chua and Asur developed two topic models for Twitter event identification [70]. They selected a set of tweets that best described each observed occurrence, integrating the time component of the tweets' keywords with the words themselves.

There are two main kinds of strategies used in the past to use social media feeds to find and summarise sporting events: graph-based methods and rate-based methods. One of the first to investigate the possibility of a sports event summary was Nichols et al. [71]. Their method is classified as rate-based and detects a sub-event when the tweet volume exceeds a certain level. They then chose illustrative tweets that used the phrase graph method to describe the various match-related sub-events. A sub-event occurrence was noted when the twitter stream rate surpassed 90% of the previously measured rate, according to Zubiaga et al. [72]. Once a sub-event had been discovered, they employed the Kullback-Leibler Divergence technique for word weighting to choose

[2] https://automatedinsights.com

[3] https://narrativescience.com

the tweets that had the highest scores for encapsulating the sub-events. Marcus et al. [73] introduced the "Twitinfo" event detection and presentation system. Using a peak detection algorithm, Twitinfo recognises sub-events based on given keywords and then presents an event timeline. The problem of delivering real-time summaries of sporting events was handled by Kubo et al. [74] by leveraging Twitter users who are recognised as great reporters. The user scores are calculated by giving higher points to users who post more often during previously identified sub-events within the event in order to identify these reporters. A real-time strategy for identifying sub-events that take place during soccer matches was proposed by Hsieh et al. [75]. They devised a two-step approach, using their moving-threshold technique to recognise sub-events and the TF-IDF to determine the most representative keywords for each sub-event. Ranking data has also been utilised for content pinpointing [76]. Aloufi et al. [77] recently proposed a framework for chronological multi-view multi-modal summarization using microblog streams for sports events.

2.4 Soccer Game Summarization

A basic framework for sports video summarization and its application to soccer footage is provided by Li et al. [78]. Baseball and American football are examples of sports that comes under action-and-pause sports in contrast to continuous activity sports like soccer. By considering each pitch as a significant event, keyframes are retrieved in the action-and-pause sports category. Following that, the clip is separated into events and non-events. In the continuous action sports case, the broadcaster's audio is analysed to identify any keyframe-worthy segments. An application for soccer is used to demonstrate the framework. Close-up shots and audible excitement are used to signal the start of an exciting event. A hybrid video summarization technique for cricket footage was developed by Javed et al. [79]. The duration of the game makes it challenging to produce cricket video recaps. In order to gauge the volume of comments and audience applause, audio data are divided into short frames and the pitch of each frame is determined. Then, in order to pinpoint the pivotal times in cricket, the framework is provided with the proper video frames of the thrilling audio samples. Boundary, six, wicket, and replay events are categorised using a decision tree and a set of event rules [80]. According to a knowledge-enhanced summarizer created by Wang et al. [81], sports news is produced by combining live commentary and the knowledgebase.

3 METHODOLOGY

Our methodology consists of utilizing soccer game audio streams (commentary and game audio) and metadata to identify and locate important events, which are then presented using customized versions of state-of-the-art journalistic article templates. It uses a variety of cutting-edge deep learning methods to generate text from audio and revised text from text input. Below, some of these approaches are explained in detail.

3.1 Background Concepts

3.1.1 Importance of Transformers in NLG

Recurrent Neural Network (RNN)s, can be visualized as a series of cells and were used intensively on encoders and decoders, one of most successful architectures in NLP. The encoder RNN receives a sentence as input and reads it token by token. Until all of the words in the sentence have been processed, each cell takes an input word and produces a hidden state as an output, which is then supplied as input to the next RNN cell.

This assumption that the last-generated hidden state reflects the essence of all the information included in each word of the input phrase is factually inaccurate in that it is more challenging when there are long sentences with several clauses. Most significantly, vanishing gradient issues affected plain vanilla RNNs. By using a forget gate, Long short-term memory (LSTM) was able to solve the vanishing gradient problem and select which information should be taken seriously and which may be disregarded. The intricate gated architecture of the LSTM somewhat resolves the issue of long-term dependency. However, as a recurrent architecture, the utilisation of parallel computation was still hindered, making LSTMs particularly sluggish to train. The LSTM must still be trained sequentially implying that dependencies might move from left to right, rather than in both ways as with the newer attention mechanisms. Attention to subset of the text was as a solution to the bottleneck issue, mimicking a human translator would pay special attention to a single term. The transformer was created as a result of the realisation that the attention mechanism alone could provide greater accuracy than the recurrent architecture with sequential word input. Transformers with Self-Attention now dominate the NLP/Natural Language Understanding (NLU), allowing for the parallel construction of the transformer architecture and enabling for training on high-performance hardware like Graphics processing unit (GPU)s, which facilitates the training of ever-larger models.

3.1.2 Basics of Transformers

The original Transformer architecture consisted of encoder and decoder blocks. An encoder block accepts inputs up to a particular maximum sequence length (e.g., 512 token, in the original transformer paper). For an input sequence shorter than the encoder input limit; the remainder of the sequence are padded.

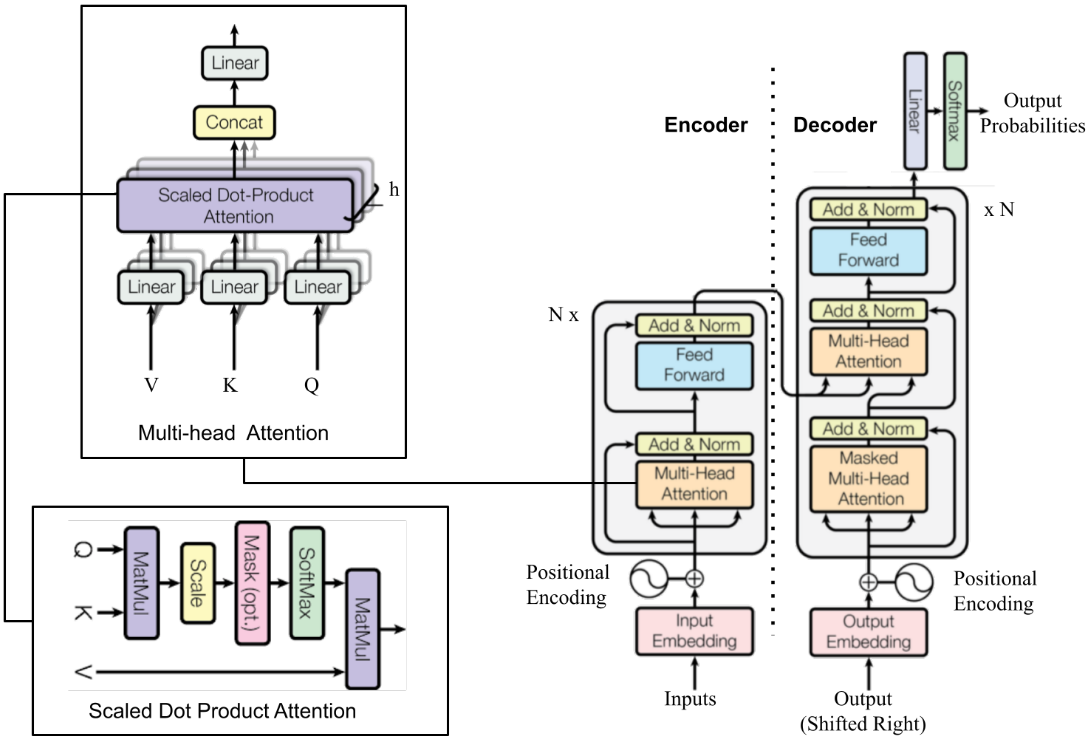


Figure 3.1: Transformer Model Architecture, Adopted from [82]

To enable it to focus on certain encoder segments, the decoder block differs somewhat architecturally from the encoder block. As opposed to Bidirectional Encoder Representations (BERT), which transformed the future words to [mask] token, the self-attention calculation blocks information from tokens to the right of the location being computed. A position may peep at tokens to its right while using normal self-attention, but when using masked self-attention, as in GPT-2, this is not possible. In Figure 3.2 it is shown that while calculating the attention, for instance for position 4, only the present and previous tokens are considered.

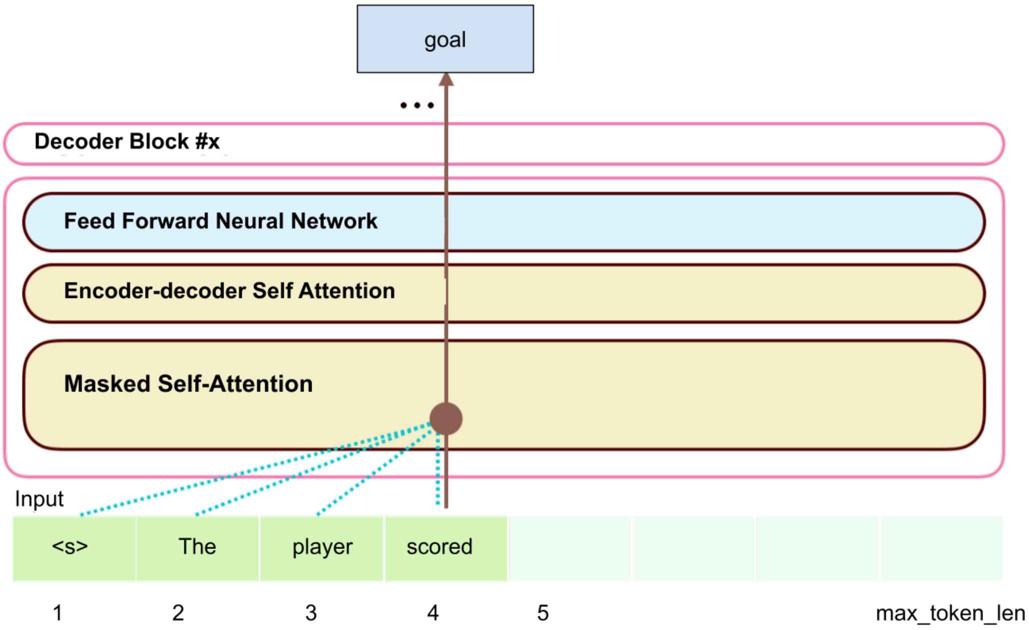


Figure 3.2: Masked Self-attention in Transformer Decoder Block

3.1.3 LongFormer

Due to the extensive matrix operations needed in the self-attention operation, transformers are costly. The transformer-based models are unable to analyse long sequences because the self-attention operation quadratically expands with sequence length. Normally, during self attention each token interacts with every other token in the input, requiring operational complexity of $O(n^2)$ for n tokens for each input sequence of length n as shown in Figure 3.3(a).

Regular self-attention is modified by Longformer's attention mechanism, which combines local windowed attention with task-driven global attention. This kind of linearly scalable ($O(n)$) sparse attention approach makes it easier to analyse texts with thousands or more tokens.

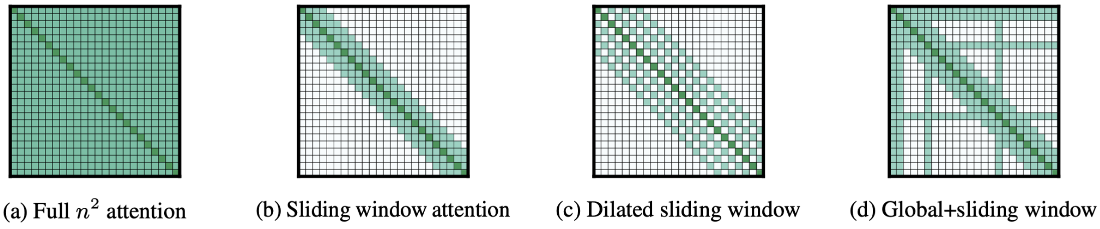

(a) Full n^2 attention
(b) Sliding window attention
(c) Dilated sliding window
(d) Global+sliding window

Figure 3.3: Attention Strategies for Transformers

Attention Patterns

Longformer utilises multi-head attention mechanism and each head with multi-head attention calculates a distinct score. Different sliding window attention mechanisms are utilised. Standard sliding window attention is used for the bottom levels and dilated sliding window attention for few upper layers to provide a satisfactory representation of all tokens. The rationale behind this method is that the local context is more significant in the lower levels, while the global context is more relevant in the top layers.

Sliding Window Attention

Similar to CNN kernels that perform a matrix operation to a set of pixels and then proceed to the next set, sliding window attention approach only pay attention to tokens in the current window altering the attention goal such that it only focuses on tokens that appear in a context window w as in the Figure 3.3(b). Each token will only be able to interact with the $\frac{1}{w}$ tokens to its left and right.

However, this restrict the amount of tokens taken into consideration to just those inside the window. To mitigate such situation, multiple layers of self-attention is used such that after successive self-attentions in multiple layers, the tokens can virtually attend other tokens not in their window in the first layer as shown in Figure 3.4 for the token 'scored'. Although the attention window of three in the first layer is only able to make it attend with two other tokens in the sequence, it gets opportunity to attend with more neighbouring tokens in the second layer. With this phenomenon, a long sequence of tokens can be attended as the depth increases.

Dilated Sliding Window

Instead of examining all tokens in window w, alternative tokens are only considered as shown in Figure 3.3(c) to tackle lengthy sequences. Skipping tokens obviously result in information loss in the lower levels, which is transmitted to the upper layers resulting in poor performance from the models as well as unstable training. However, this little modification accommodates a broader length of tokens without requiring costly architectural modifications and by widening the scope of the sliding attention window.

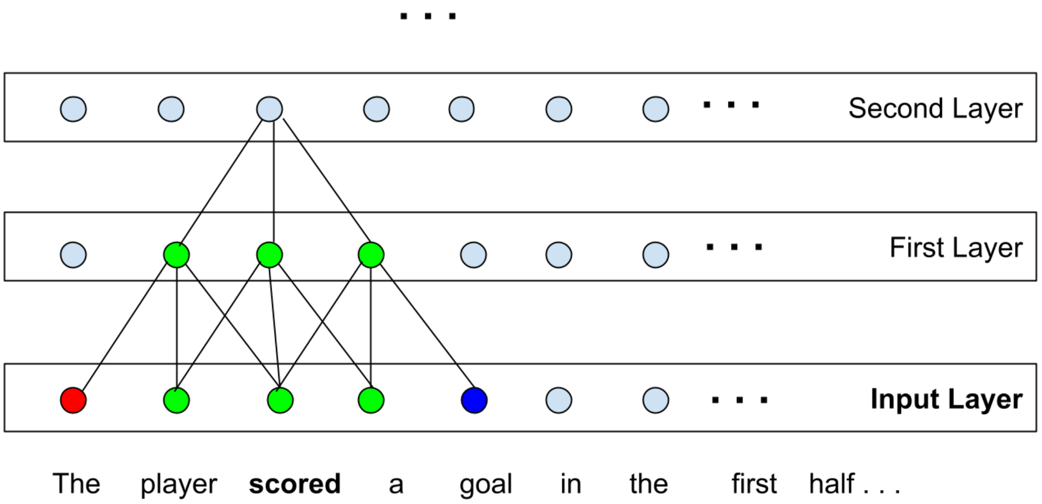


Figure 3.4: Multiple Layers of Self-attention.

Global Attention

Attentions that are windowed or dilated are insufficiently flexible to learn representations for downstream tasks. Certain tokens are allowed to perform global attention, meaning that all tokens in the sequence may attend to these tokens as shown in Figure 3.3(d).

The addition of global attention increases the performance as well as also boosts the model's representational capacity. However, due to the restricted amount of tokens, the complexity remains ($O(n)$). Configuring global attention tokens is an engineering choice and may largely influence the performance.

Linear Projections

Two distinct sets of Q, K, and V matrices will be used for sliding window and global attention. The design decisions for length of sliding window as well as global attention strike a compromise between performance and efficiency. Lower size of attention window enables quicker calculation and larger attention window allows for greater representational power.

3.1.4 Generative vs. Regressive Language Models

The OpenAI GPT-2 model addresses up to 4,000 input tokens instead of the original transformer's 512 by using Decoder-only blocks and omitting the Transformer encoder. In contrast to this, BERT uses transformer encoder blocks to learn an encoded representation of the inputs by corrupting them to recreate the original versions, thus referred to as an "autoencoder". Another key difference between GPT2 and BERT is that it produces one token at a time, similar to classical language models. After each token is generated from a well-trained GPT-2, it is added to the sequence of inputs in the subsequent phase. This approach is known as "auto-regression," which was also one of the ideas that gave RNNs their disproportionate success. The nature of certain subsequent models, including GPT-2 and TransformerXL, is auto-regressive. But it is clear that even if BERT did away with auto-regression, it might still achieve better results if it could make use of the context of both sides of a word. In this way, XLNet reintroduces autoregression while including context on both sides in a novel manner: this is a kind of compromise.

BERT is trained to map latent relationships between the text of various contexts, sentiment analysis and question responding are predicted based on a deep bidirectional context. However, the GPT-3 training technique is rather straightforward in comparison to BERT and is favored when little data is provided, with a wider variety of applications. It has been shown that a sufficiently complicated network that incorporates the context of the posterior sequence of words when selecting a word suffers knowledge leakage. To address this, BERT does some % of token masking to prevents cheating via rote memory by withholding just enough information. Because a token at a certain place in a phrase only has access to prior tokens, the GPT-2 training is also a natural fit for constrained summarization as it learns directly from the "predict next" task.

3.1.5 Self-Attention in Transformers

The input transformer block takes a input of $\mathbf{x} \in \mathbb{R}^{n \times d}$ length- n sequence of d-vectors of template soccer summary. A transformer's L transformer blocks are each parameterized function classes, $f_\theta : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d}$, and have their own set of parameters: $f_{\theta_L} \circ \cdots \circ f_{\theta_1}(\mathbf{x}) \in \mathbb{R}^{n \times d}$, implying that the output $\mathbf{z} \in \mathbb{R}^{n \times d}$ from each block has the same dimension as the input. If "queries," "keys," and "values," respectively, are represented by the variables Q, K, and V, then for each attention head in the block from $h = 1, ..., H$, W_h is the trainable weight matrix such that $W_{h,q}, W_{h,k}, W_{h,v} \in \mathbb{R}^{d \times k}$, then for

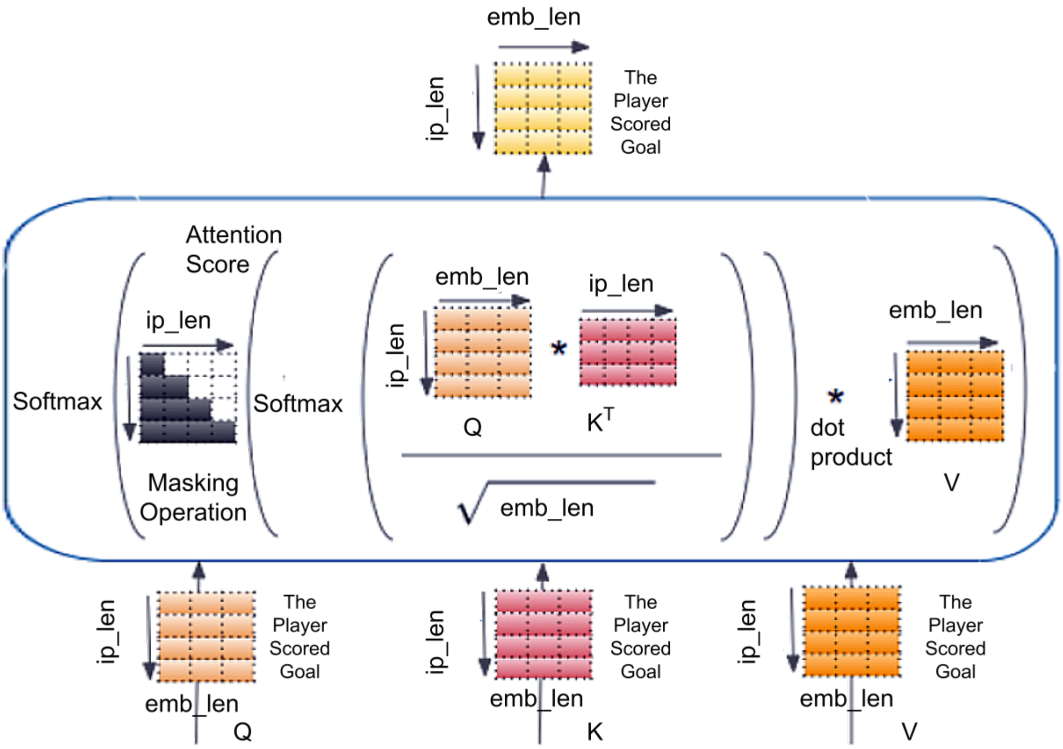


Figure 3.5: Decoder Self-Attention and Masking

each set of attention head:

$$Q^{(h)}(\mathbf{x}_i) = W_{h,q}^T\mathbf{x}_i, \quad K^{(h)}(\mathbf{x}_i) = W_{h,k}^T\mathbf{x}_i, \quad V^{(h)}(\mathbf{x}_i) = W_{h,v}^T\mathbf{x}_i$$
$$\alpha_{i,j}^{(h)} = \text{softmax}_j\left(\frac{\left\langle Q^{(h)}(\mathbf{x}_i), K^{(h)}(\mathbf{x}_j)\right\rangle}{\sqrt{k}}\right) \tag{3.1}$$

The amount that element x_i "attends" x_j in a certain head h is determined by the attention weights $\alpha_{i,j}$ in Equation 3.1 [82]. This is **self-attention** since x's components pay attention to one another and also lack explicit learning parameters during training.

softmax$_j$ denotes the application of softmax to the d-dimensional vector indexed by j.

Positional Encoding: Transformer sees a collection of n unordered, d-dimensional input vectors as a bag of features without the concept of positional encoding. The input is unordered in the sense that it is unaware of the relationship between its n inputs, denoted by $u_i \in \mathbb{R}^d$. To represent positions in a transformer, the positional connections must be expressed as data. The simplest method involves encoding locations as 1-hot features. Assume that the information in $\mathbf{x} \in \mathbb{R}^{n \times d}$ is organised progressively down the n-dimensional axis. The k'th standard basis vector in $\mathbb{R}^n$ is represented by $\mathbf{x}$, which is then extended to a new series $\mathbf{x}' \in \mathbb{R}^{n \times (n+d)}$, where $\mathbf{x}_i' = (\mathbf{x}_i, \mathbf{e}_i)$. Knowledge of the combined representation z $\in \mathbb{R}^{n \times d}$ can be specified by:

$$\mathbf{z}_k = W_z^T \operatorname{ReLU}\left(W_x^T \mathbf{x}_k + W_e^T \mathbf{e}_k\right) \in \mathbb{R}^d, \quad W_x \in \mathbb{R}^{d \times m}, W_e \in \mathbb{R}^{n \times m}, W_z \in \mathbb{R}^{m \times d}$$

Gehring et al. [83] create distinctive representations of inputs and positions:

$$\mathbf{z}_k = W_{zx}^T \operatorname{ReLU}\left(W_x^T \mathbf{x}_k\right) + W_{ze}^T \operatorname{ReLU}\left(W_e^T \mathbf{e}_k\right) \quad W_x \in \mathbb{R}^{\dim(x) \times m}, W_e \in \mathbb{R}^{n \times m}, W_{zx}, W_{ze} \in \mathbb{R}^{m \times d}.$$

In Vaswani et al. [82], sinusoidal position embeddings were used, $\mathbf{p} \in \mathbb{R}^{n \times d}$, to generate fixed representation of positions, which functions similarly to a learned one and even generalizes well to longer sequences.

$$\begin{gathered}\mathbf{p}_{k,2i} = \sin\left(\frac{k}{10000^{2i/d}}\right), \quad \mathbf{p}_{k,2i+1} = \cos\left(\frac{k}{10000^{2i/d}}\right) \\ \mathbf{z} = W_{zx}^T \operatorname{ReLU}\left(W_x^T \mathbf{x}_1\right) + \mathbf{p}.\end{gathered} \tag{3.2}$$

3.1.6 Fine-tuning GPT

Given the context vectors of input tokens of template game summary, $U = (u_1, \ldots, u_n)$ the embedding matrix We and positional encoding matrix W_p is applied to U in order to get the input vector for the decoders at the bottom level: $h_0 = UW_e + W_p$. The vector h is feed through the decoding stages, n being the number of decoders in the stack of decoders:

$$h_l = \text{transformerblock}\left(h_{l-1}\right) \forall i \in [1, n] \tag{3.3}$$

The transformer block make use of masked self-attention. The vector output from the final decoder as in Equation 3.3 is transformed into the final probability vector output: $P(u) = \operatorname{softmax}\left(h_n W_e^T\right)$. The maximization objective is as in Equation 3.4:

$$L_1(U) = \Sigma_i \log P\left(u_i \mid u_{i-k}, \ldots, u_{i-1}; \Theta\right). \tag{3.4}$$

The labelled dataset C consisting of a series of input tokens of game summary $x^1, \ldots, x^m$, and corresponding expected summary y, does not affect every model parameter. Instead, it goes through our pre-trained model and receives the final transformer block activation h_l^m, which is then supplied to an additional linear output layer with parameters W_y to forecast each token for output y based on the conditional probability: $P\left(y \mid x^1, \ldots, x^m\right) = \operatorname{softmax}\left(h_l^m W_y\right)$. The objective then becomes:

$$L_2(C) = \Sigma_{(x,y)} \log P\left(y \mid x^1, \ldots, x^m\right). \tag{3.5}$$

Including language modelling as an auxiliary target to this fine-tuning improves generalization and speeds convergence of the objective in Equation 3.5, where λ being a weight factor. The improved objective is shown in Equation 3.6:

$$L_3(C) = L_2(C) + \lambda L_1(C) \tag{3.6}$$

The fine-tuning step are then not as expensive as pre-training, as only the embedding and the newly added linear output layer W_y are fine tuned.

3.1.7 Root-mean-square (RMS) Analysis

Using the audio samples or spectrogram, the RMS value for each audio frame may be determined. A signal's RMS value is calculated as the square root of the average of the squared values of the signal sample.[84]. For a collection of complex-valued signals represented as N discrete sampled values $-[x_0, x_1, \cdots, x_{N-1}]$, the mean square value x_{RMS} is provided as in Equation 3.7 [84].

$$x_{RMS} = \sqrt{\frac{|x_0|^2 + |x_1|^2 + \cdots + |x_{N-1}|^2}{N}} = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} |x[n]|^2} \tag{3.7}$$

RMS value calculation from audio samples seems to be efficient since no STFT computation is required. Since spectrogram frames can be windowed, spectrograms provide a more precise depiction of energy across time; hence, the use of spectrogram for RMS value calculation is preferred. Using frequency domain components $X[k]$ and Parseval's theorem, it is also possible to calculate the root mean square value as in Equation 3.8.

$$x_{RMS} : \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} |x[n]|^2} = \sqrt{\frac{1}{N^2} \sum_{k=0}^{N-1} |X[k]|^2} \tag{3.8}$$

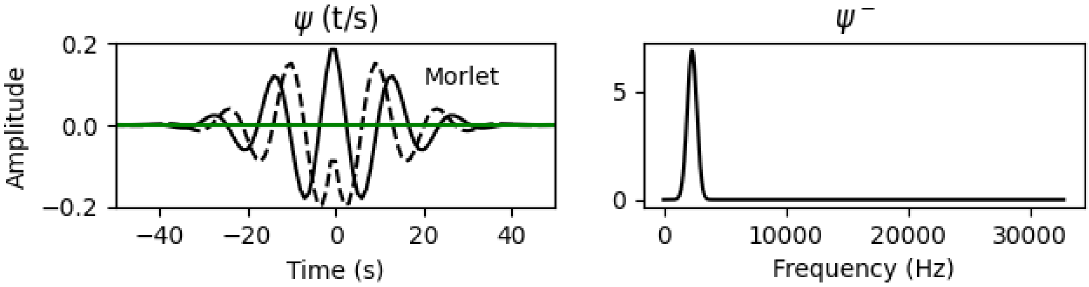


Figure 3.6: Mother Morlet Wavelet

3.1.8 Wavelet Analysis

The wavelet has the benefit of combining a wave with a certain period and being limited in size. The Morlet wavelet, a popular and straightforward wavelet, as seen in Figure 3.6, is actually a Sine wave multiplied by a Gaussian envelope.

The formula for mother wavelet for the Morlet wavelet transform is shown in Equation 3.9 [85].

$$\psi_0(\eta) = \pi^{-1/4} e^{i\omega_0 \eta} e^{-\eta^2/2} \tag{3.9}$$

The number of oscillations inside the wavelet itself is denoted by the wavenumber w_0. In practice, $w_0 = 6$ is set such that the wavelet's mean is zero. To effectively sample all frequencies included in the time series, a set of scaling parameters s, also known as the scale or period, is selected. First, the lowest resolvable scale, $s0$, is selected and then multiplied by a constant multiple up to the maximum scale. The greatest scale should be less than half of the whole time series.

3.2 Datasets

This section describes and discusses the existing datasets: SoccerNet, HOST, SportsSum and K-SportsSum Dataset, that are going to be used for the thesis. Some of these are also used to create training and testing sets for the model training and finetuning.

3.2.1 SoccerNet Dataset

SoccerNet [86] consists of a total of 764 hours from 500 different untrimmed broadcast soccer games annotated with three primary events classes: goal, yellow/red card, and player substitution.The coarse data, which is automatically derived from event reports, is manually adjusted to a precision of one second before being used. The localisation of sparse events inside lengthy game videos is the primary goal of this dataset.

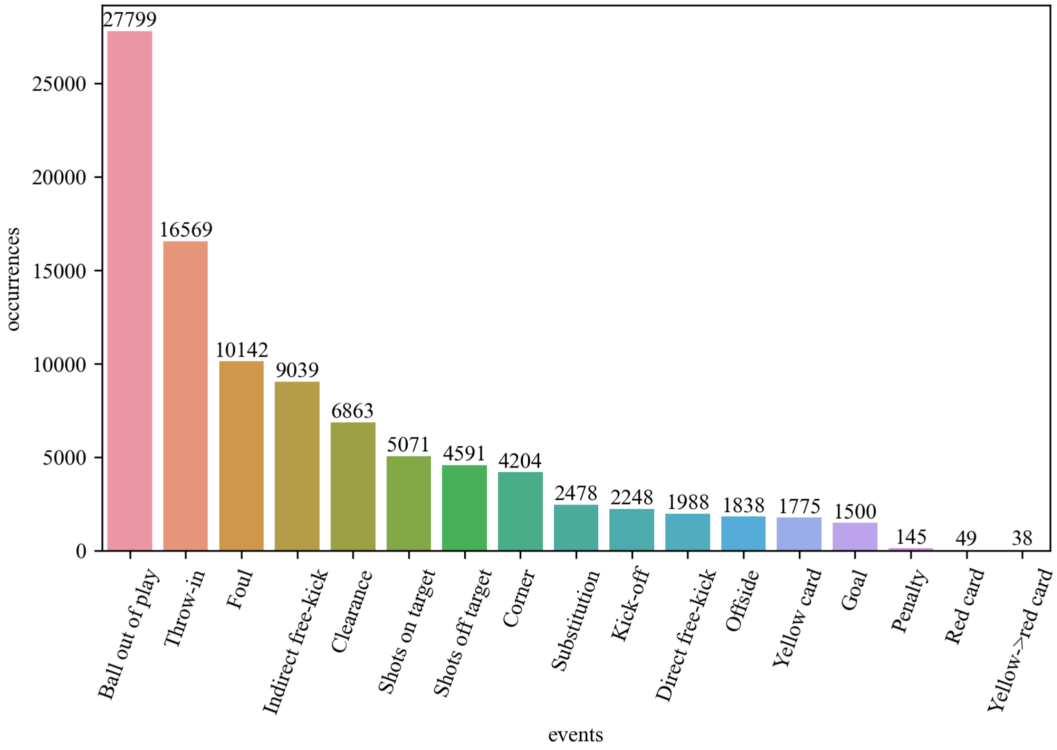


Figure 3.7: Distribution of Events in SoccerNet Dataset Meta-data

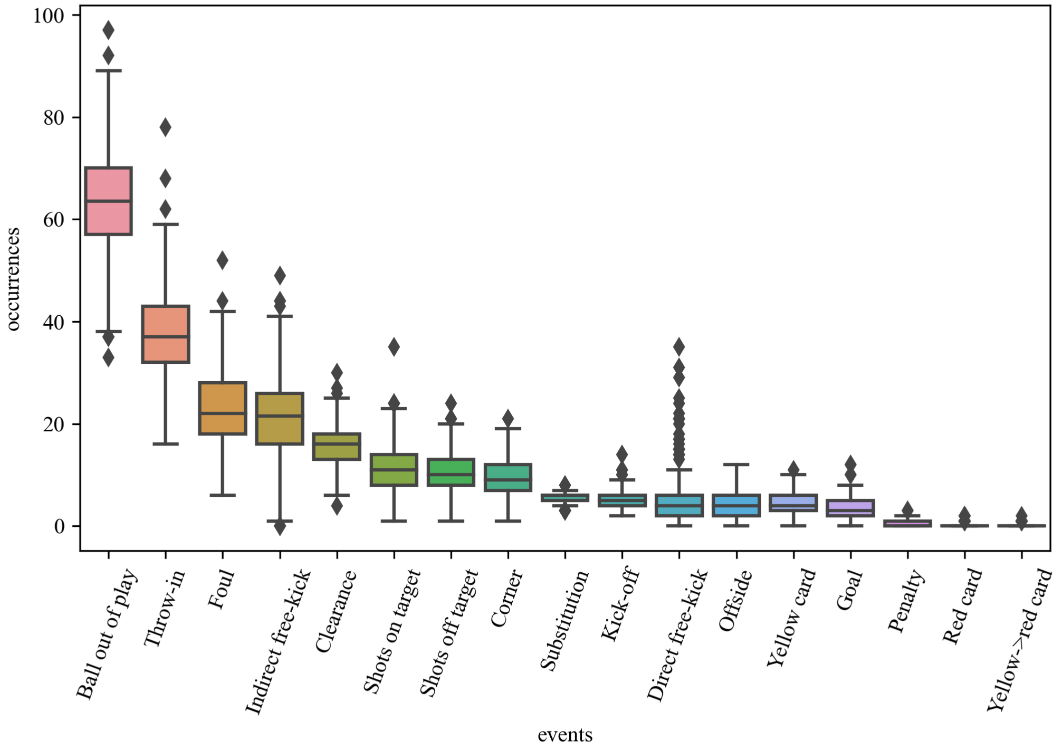


Figure 3.8: Per-game Distribution of Events in SoccerNet Dataset Meta-data

SoccerNet-V2 [87] extends the action classes from three to seventeen to support event spotting on soccer videos. It also emphasizes temporal segmentation of camera shots and retrieval of the replayed actions in the game. The structure of the dataset for the action spotting task of SoccerNet-V2 is shown in (Listing A.2) along with event annotation structure in (Listing A.1). Distribution of Events in SoccerNet dataset is shown in Figure 3.7. Likewise, distribution of events per-game is shown in Figure 3.8.

3.2.2 HOST Dataset

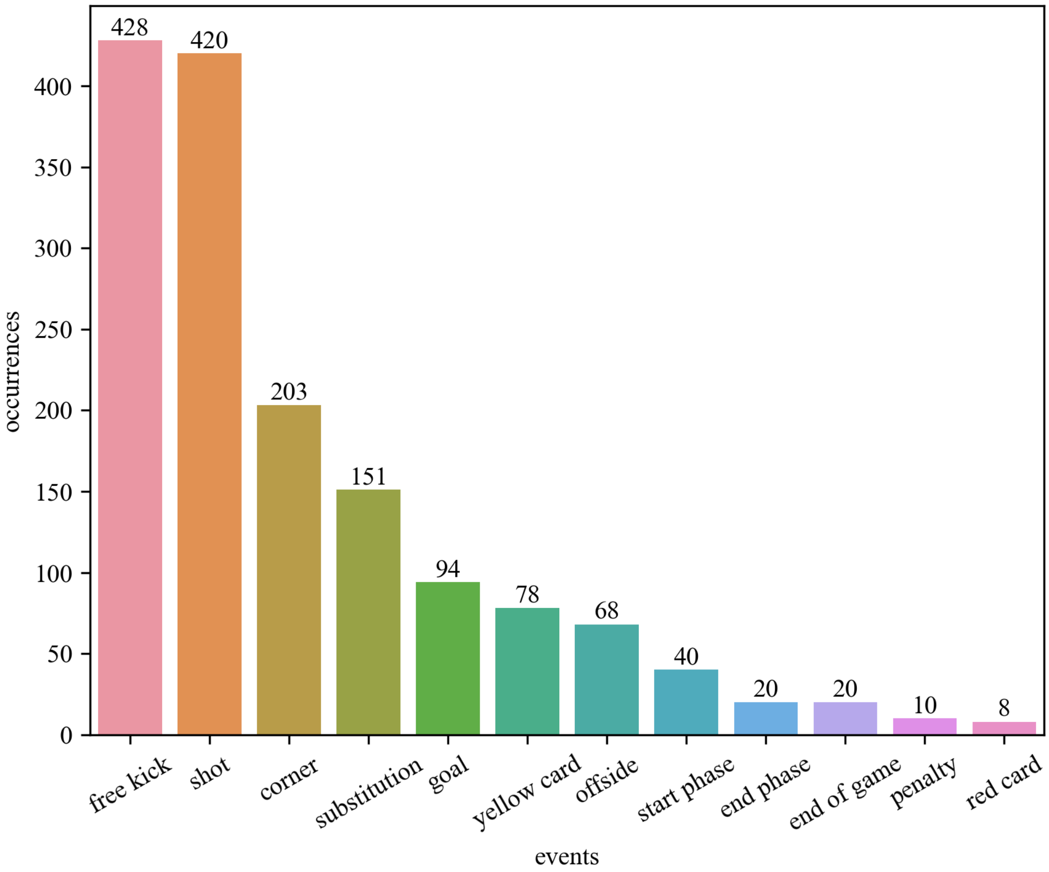


Figure 3.9: Distribution of Events in HOST Dataset Meta-data

This dataset comprises of 15 full soccer game videos from the Norwegian *Eliteserien*, with an event annotation list of each game's highlights. Highlights include extra timing meta-data(Listing A.4)[1], goal annotations (Listing A.3), and card annotations (Listing A.5). Employees at ForzaSys AS and researchers from the Simula Metropolitan Center for Digital Engineering(SimulaMet) curated the dataset.

[1] Starting with the line, each meta-data list reads:"Video start timestamp: <YYYY-MM-DD HH:mm:ss.ssssss>".

Distribution of Events in SoccerNet dataset is shown in Figure 3.9. Likewise, distribution of events per-game is shown in Figure 3.10.

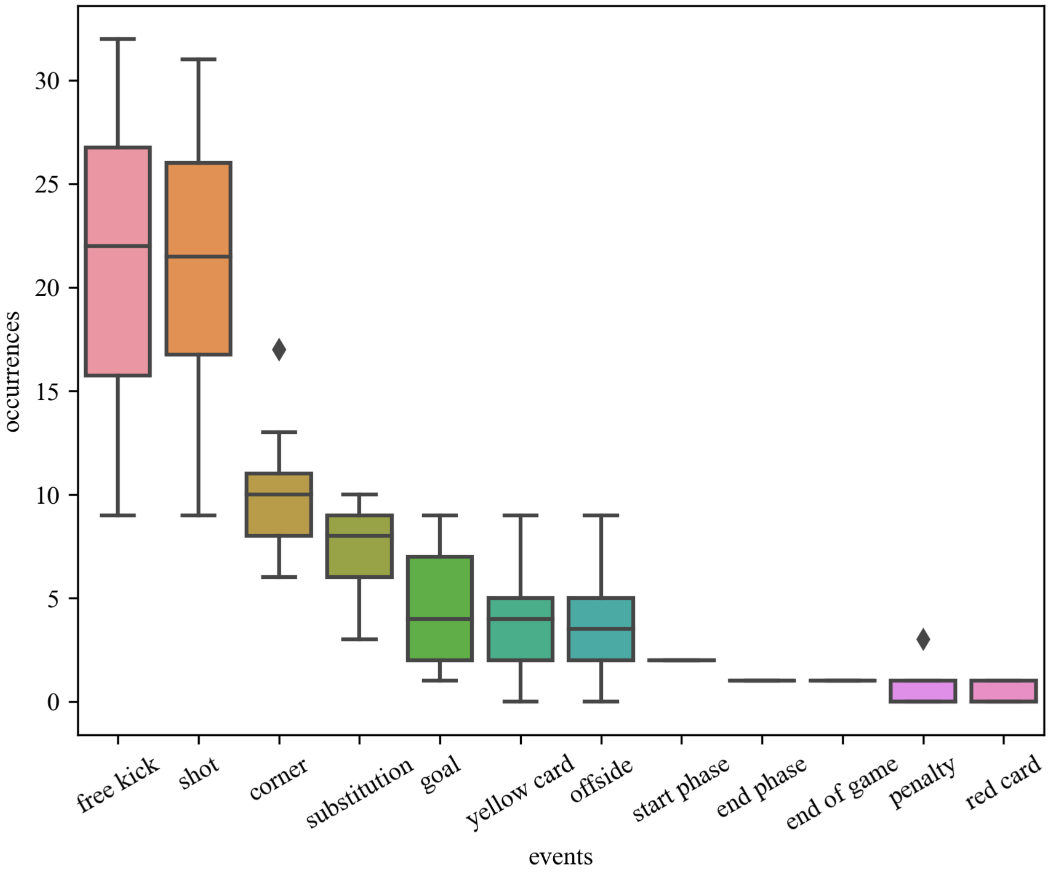


Figure 3.10: Per-game Distribution of Events in HOST Dataset Meta-data

3.2.3 SportsSum Dataset

SportsSum [88] has a total of 5428 soccer game commentaries with corresponding news scrapped from online sources in Chinese.

3.2.4 K-SportsSum Dataset

The K-SportsSu [81] dataset includes summaries of 7854 worth of sports matches along with a large knowledge corpus that includes data on 523 sports teams and 14K+ worth of sports players in Chinese. In order to increase the dataset's size and quality, the original authors de-noised news data using a comprehensive manual cleaning process so that the news's scope was restricted to the specific game.

3.3 Proposed Framework

3.3.1 Components

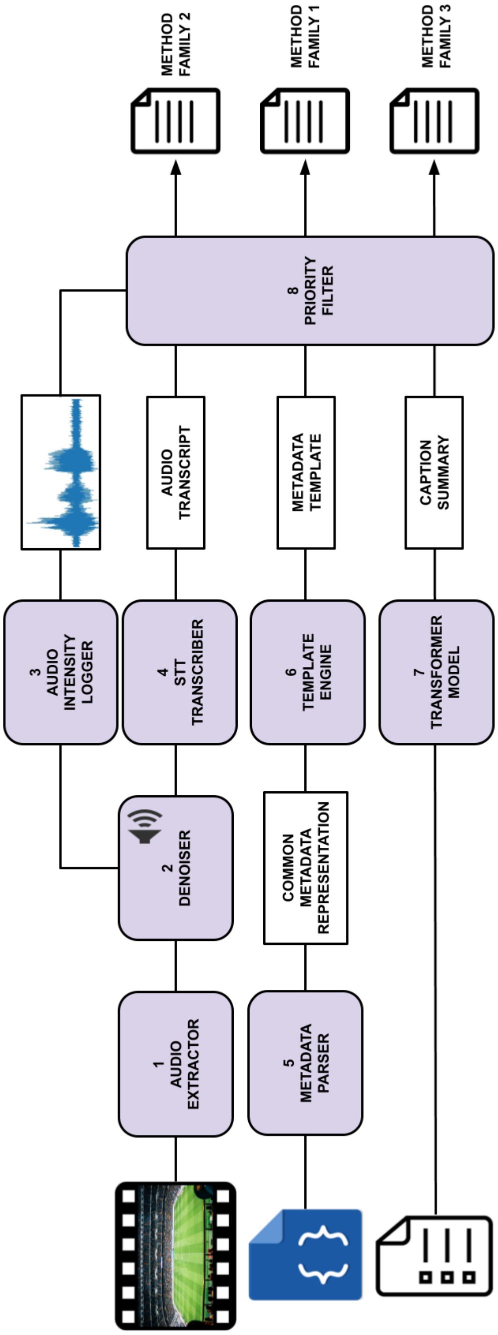


Figure 3.11: Block Diagram of the System

Figure 3.11 presents the system block diagram of our proposed framework. The end-to-end pipeline comprises 8 main modules. The key blocks in the diagram are explained below.

Module 1: Audio extraction

This module is for extracting an audio stream from the video stream. Fast Forward Moving Picture Experts Group (FFMPEG) [89], a popular software framework for transcoding multimedia files including audio and video is used. Popular video formats include Moving Picture Expert Group-4 (MP4), QuickTime Movie (MOV), (Audio Video Interleaved (AVI), Flash Video (FLV), and Matroska Video (MKV). It handles anything from the oldest and most esoteric formats to the newest and most up-to-date ones. The audio bitrate is set to 128k and the audio sampling frequency to 44100Hz. This particular configuration is used for being the default configuration for Spleeter. The pre-trained audio separation model has been trained with audios with a sampling frequency of 44100Hz.

Module 2: Background/noise removal

This module is for de-noising the extracted audio. As mentioned above, Deezer spleeter [90] is used for this purpose. It outputs separated vocals and accompaniment files. The vocal audio file contains commentary audio with filtered background noises. Note that any other industry standard noise suppression mechanism can be used instead of Spleeter in this module. Depending on the presence of commentary audio in the video, noise suppression can be omitted (e.g., for videos where commentary is not present and the audio intensity during the game is the only aspect of interest).

Module 3: Audio intensity log generation

This module records the audio intensity, which is later used to filter the temporally annotated metadata or captions. The time information corresponding to a certain level of audio energy is used to pinpoint the exact timeframe in the video and the extracted audio. A configurable number of audio frames are analyzed to retrieve the intensity level. As rising audio intensity levels generally follow important highlights (aftermath), the selected set of frames mostly contain audio levels after the event has occurred, with an additional few frames from before the event also included. Figure 3.12 demonstrates two alternative methods of calculating audio intensity, namely RMS energy and zero crossing.

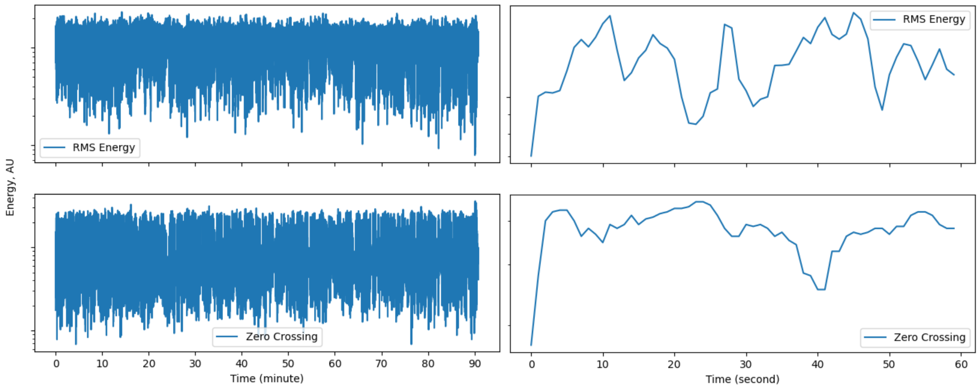


Figure 3.12: Audio Intensity Levels for a Sample Game Using RMS and Zero Crossing. Left: Complete Game (90 Minutes), Right: Zoom In on First Minute (60 Seconds).

As part of the pipeline, an easy-to-use dashboard is implemented for understanding the correlation between audio intensity levels and the events in the game video. Figure 4.3 presents a screenshot from the audio intensity analysis dashboard, which plays the game video along with indicators for the corresponding audio levels, event annotations, and an ordered list of the top events in the game during which the audio intensity was highest. The dashboard can be used as a validator for the filtering step (Module 8). This tool is provided as an open source software[2] for the community.

Module 4: Audio transcription using STT

In this module, the noise-free audio commentary is processed by a speech recognition system to convert the audio into text. Alternative third party tools which can be used for this purpose include Amazon Transcribe [91], Azure Speech to Text [92], Google Speech to Text [93], and IBM Watson Speech to Text [94]. System capabilities such as language support make up an important aspect for generalizability (e.g., the transcription of leagues from non English speaking countries in Scandinavia, such as the Norwegian Eliteserien or the Swedish Allsvenskan, was not possible using some of the above alternatives), and also present a trade-off between scale and cost. In the current implementation, IBM Watson is used in this module.

[2] https://github.com/simula/soccer-summarization

Module 5: Metadata parsing

Different soccer datasets can include game metadata in different formats and use different annotation styles. For uniformity, metadata need to be parsed and translated into a standard format[3] . This module currently includes support for 3 different metadata input types (in-house HOST dataset, SoccerNet, and K-SportsSum), which can be translated into a common metadata representation.

Module 6: Template generation

In this module, templates are applied to the metadata to generate static summaries. The knowledge base contains event prioritization rules and other domain expert know-how which can be leveraged in training the system. Table 3.1 presents examples of sentence generation using a naive template, for two different datasets.

Module 7: Natural language generation with transformers

Language models such as transformers can be used to distil commentary texts into news. Self-attention, which is the building block of transformers, being a costly operation, limits the total number of words that can be fed to or can be expected as output from the transformers.

New models, such as Longformers [96], substitute a task-motivated global attention approach for the traditional self-attention operation via local windowed attention. Such a mechanism for self-attention allows a large number of tokens as the input, making it a good candidate for summarization of large input texts. Models with 16384 tokens as input and 1024 tokens as output have shown promising performance on multiple downstream tasks, including summarization. The model can be finetuned to generate news from long commentary texts.

[3] In cases where metadata are not available, and only game commentaries or captions are available, trained language models can be used to construct metadata. Such a process can help enrich the information available, as well as filter out irrelevant or unimportant content. A finetuned GPT-3 [95] model has been shown to be a good candidate for generating metadata directly from machine generated captions and human commentary.

Table 3.1: Template for the Generation of Naive Interpretations From Metadata.

Dataset	**Metadata Format**	**Interpretation**	**Sample Sentence**
HOST	('free_kick', 'offending_player', 'team')	d[team][value] was awarded a free kick because of d[offending_player][value].	Bodø/Glimt was awarded a free kick because of Erling Haaland.
SoccerNet	('free_kick', 'team')	d[team][value] was awarded a free kick.	Bodø/Glimt was awarded a free kick.
HOST	('red_card', 'player', 'team')	d[player][value] from d[team][value] got a red card.	Sondre Sørli from Bodø/Glimt got a red card.
SoccerNet	('red_card', 'team')	d[player][value] got a red card.	Bodø/Glimt got a red card.
HOST	('goal', 'assist_by', 'scorer', 'shot_type', 'team')	d[scorer][value] scored a goal by d[shot_type][value] shot for d[team][value] with assistance from d[assist_by][value].	Sondre Sørli scored a goal by right-footed shot for Bodø/Glimt with assistance from Japhet Sery.
SoccerNet	('goal', 'team')	d[team][value] scored a goal.	Bodø/Glimt scored a goal.
HOST	('substitution', 'player_in', 'player_out', 'team)	d[player_in][value] replaced d[player_out][value] in d[team][value].	Patrick Berg replaced Sondre Sørli in Bodø/Glimt.
SoccerNet	('substitution', 'team)	d[player_in][value] replaced one of its players.	Bodø/Glimt replaced one of its players.

Module 8: Priority filtering using audio intensity

In this optional module, prioritization rules are used to weigh candidate sentences or particular events in the game that would be included in the summary. Audio intensity levels as identified by Module 3 are utilized. Such a prioritization helps the overall pipeline create a maximally informative summary for a given set of conditions, such as length constraints.

3.3.2 Alternative Summarization Methods

In this section, alternative end-to-end summarization methods which can be run using suggested pipeline is described. Table 3.2 presents an overview of these methods, where two adopt naive approaches and the third is the proposed approach.

Table 3.2: Alternative Methods for End-to-end Game Summarization With/without Denoising and Audio-Based Priority Filtering (M: Metadata, A: Game Audio, C: Captions).

ID	Input	Family	Denoising	Priority Filtering
1.1	M	Naive metadata template	✗	✗
1.2	M+A		✗	✓
1.3	M+A		✓	✓
2.1	A	Naive STT	✗	✗
2.2	A		✓	✗
2.3	A		✗	✓
2.4	A		✓	✓
3.1	C	Transformer model	✗	✗
3.2	C+A		✗	✓
3.3	C+A		✓	✓

Method 1: Summary from game metadata using naive template

The game metadata (examples in Listings A.1 and A.3) along with a naive template as exemplified in Table 3.1 is used to generate text summaries. A priority mechanism based on audio intensity can be employed to explicitly filter important events as generate summaries as shown in Figure 3.11 (Method Family 1).

Method 2: Summary from game audio using naive STT transcription

STT is employed for getting text transcripts of the human commentary from the game audio. Denoiser module is used to clean the audio from unnecessary background noise before feeding to the STT engine. The scope of such texts is generally wider as the audio contains the conversation referring to the history of the team or players, the status of the team in the league, etc. A text-based filtering mechanism can be used to remove redundant sentences, and an audio-based filtering mechanism can be used to identify relevant lines of the transcribed text that are deemed important, for inclusion in the summary as illustrated in Figure 3.11 (Method Family 2).

Method 3: Summary from game commentary using transformer model

A transformer model is trained so that, for a given set of time-stamped game caption texts, it predicts the summary of the game. The input and output limits of this method are constrained by the capabilities of the transformer model used. Due to the length constrains in the transformer, the audio intensity log can be used to filter irrelevant and unnecessary sentences, or, prioritise important sentences. This method is indicated in the Figure 3.11 (Method Family 3). An additional text-based filtering module can help to filter irrelevant lines of the text in the output from the text generation module. This can also be used to filter false information. However, this module is out-of-scope of the thesis.

3.3.3 Static Template Generation

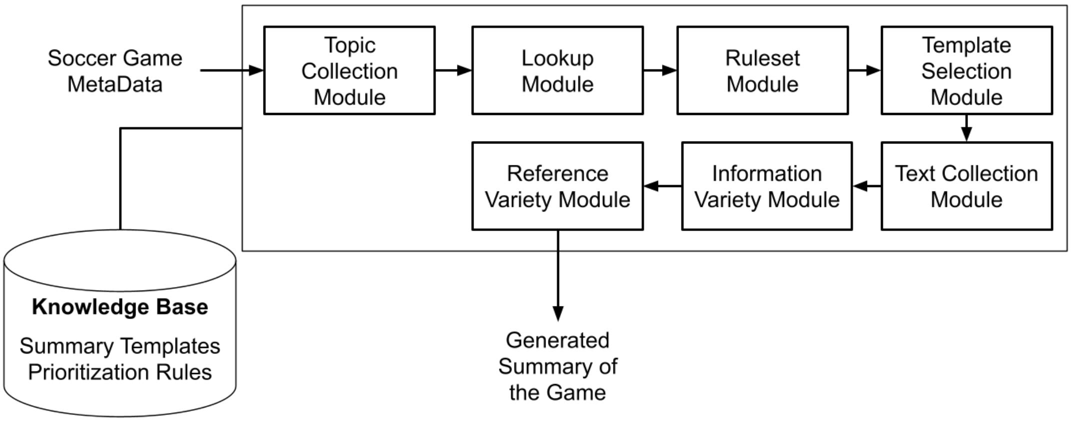


Figure 3.13: Modules for Static Template Generation

Modules for creating static templates from meta-data is shown in Figure 3.13. The topic collection module collects match-related topics and chronologically organises them. All template categories and appropriate event-description templates are obtained via the lookup module by accessing the template database. The ruleset module verifies that the conditions for using each template category are satisfied. The available templates are chosen at random by the template selection module depending on their weights. The text collection module combines the text elaborating on each instance in a predetermined order. The inclusion of an information diversity module ensures that certain types of data are not repeated in the report. The reference diversity module makes an effort to recognise the same referent when it appears twice in one speech. The module uses a different form to address the referent of the second sentence if it can recognise this.

3.3.4 Noise Suppression in Audio

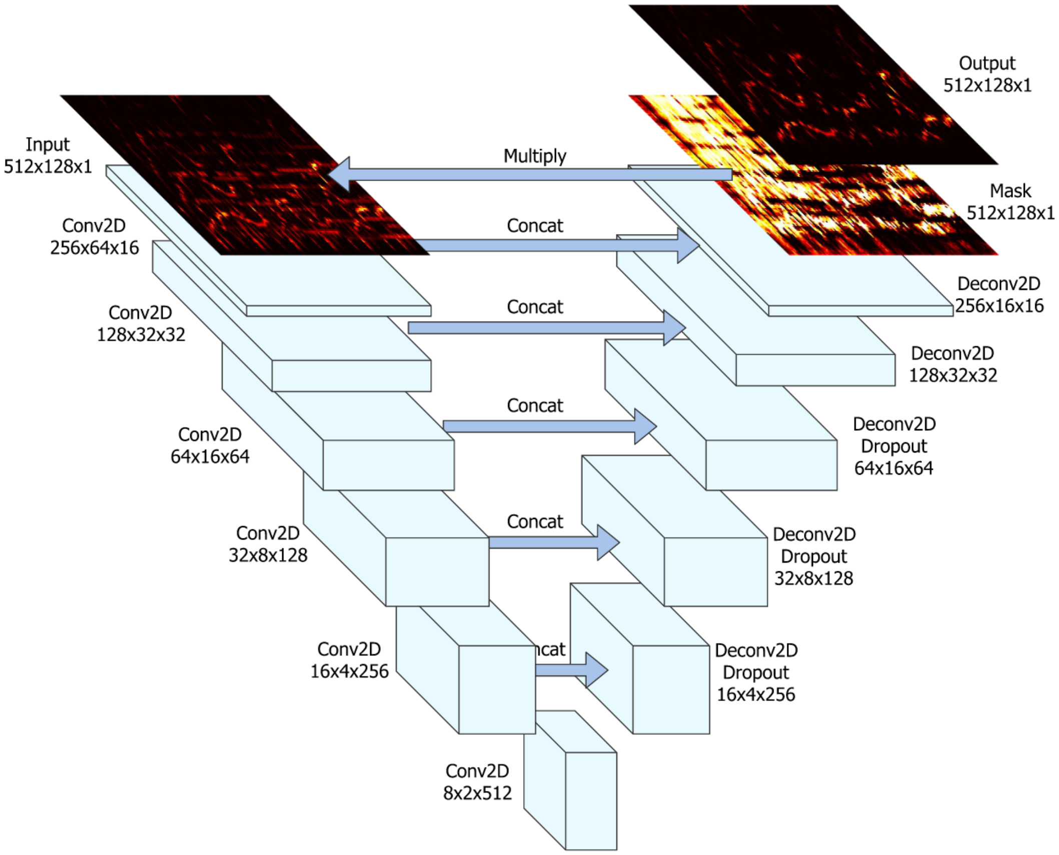


Figure 3.14: The U-Net Architecture Used in Spleeter by Deezer. Image by Rachel Bittner.

The popular source separation solution Spleeter [90] by Deezer uses U-Net as shown in figure 3.14. The network receive a spectrogram as input and execute a succession of 2D convolutions, each of which encodes a representation of the input that is progressively reduced. The compact representation is then enlarged by decoding using the same number of 2D deconvolutional layers (also termed transpose convolution), each of which matches to the form of one of the convolutional encoding layers. Each encoding layer is then combined with its respective decoding layer. Since U-Net use convolutional operation, it processed a spectrogram with a defined input size. The audio signal must be dissected into spectrograms with the same number of time and frequency dimensions used to train the U-Net. U-Net originally is a 2D convolutional network with a kernel size of 5x5 and stride size of 2. A batch norm and ReLU activation is performed after each layers. It has 6 encoder layers where on the first three encoder layers, a Dropout of 50 percent is applied. On the decoder side it has five similar convolutional layers. A mask is created by the last layer's sigmoid activation function which is multiplied by the input mixture to separate the noise from audio. The loss an be calculated simply by comparing the ground truth source spectrogram to the output spectrogram multiplied by calculated mask.

3.3.5 Audio Intensity Analysis

Audio intensity analysis can support automatic video summarization pipelines, by enabling the calculation of excitement around different events, and can be used as a filter to select game highlights based on audience-perceived importance. Alternative formulations to be used in the pipeline for audio intensity analysis and excitement calculations are described in section 3.1.7 and 3.1.8.

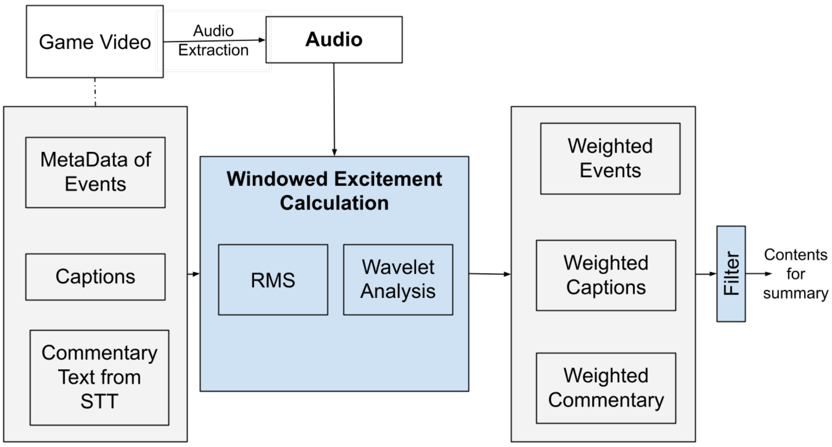


Figure 3.15: Audio-based Filter in Summarization Pipeline

Figure 3.15 shows the use of audio filter in summarization pipeline. The audio from a particular event is used to calculate excitement around that event and filter the events based on their importance.

3.4 Verification and Validation Procedures

3.4.1 User Study-based Evaluation

Subjective user validation studies are conducted in settings where people vote on alternative game summaries (text and audio) and compare them with benchmarks from literature. Also an expert study with people who have experience in the soccer (production, journalism, etc.) domains can evaluate the generated output over multiple metrics. Typically, when evaluating a text generated by an NLG system, informativeness and coherence aspects are considered.

Alternately, in a more comprehensive evaluation, evaluators are tasked with assessing fluency, informativeness, non-redundancy, referential clarity, and structure and coherence. The major aspects to be evaluated are outlined below.

Informativeness: The summary must communicate the main ideas from the incoming data. For instance, a game's key events should be included in the summary. A summary that merely mentions the goal events, nor one that is unnecessarily lengthy/verbose is not so effective. The summary should be useful from utility perspective, often referred as usefulness or helpfulness.

Non-redundancy: The summary should not repeat any ideas and, preferably, cover as much ground as possible within the word limit.

Clarity of references: References inside or between sentences in the summary should be clear and pertinent to its subject matter. For example, if a pronoun is used to refer to a player, the player should be included in the summary before the pronoun. The particular item or information (such as a prior event) being referred should also be clear and unambiguous. In addition, the language quality, often known as fluency, clarity, or readability should be evaluated from the perspective that if the text is simple to read and comprehend.

Focus: The summary must have a central theme, and each sentence must contain information pertinent to this theme. For instance, when summarizing a news about a victory with great big goal difference, the more focus could be on a major goal events and should focus less on other irrelevant information about the game.

Organization and Coherence: Not just a list of unconnected facts, the summary should be a well-organized and logical body of information. To preserve a continuous flow of information, the phrases should specifically be connected. The evaluation also should include accuracy, also known as correctness at times which addresses if the whole output text are accurate and derived from the supplied data.

3.4.2 Objective Evaluation

Human assessment has the drawback of being subjective, with a broad range of opinions on what constitutes a "good" summary. This variation suggests that developing an automated assessment and analysis approach is very complex and difficult. In automatic evaluation, ATS-generated summaries are evaluated using automated measures to lower the cost of review.The automated assessment metrics, however, still need human labour since they compare the summaries produced by the system with one or more summaries that were supplied by people.

Precision score metrics: Equation 3.10 is used to calculate precision score by dividing the number of sentences in the candidate summary by the number of sentences shared by the reference and candidate (i.e. system).

$$\text{Precision} = \frac{S_{\text{ref}} \cap S_{\text{cand}}}{S_{\text{cand}}} \tag{3.10}$$

Recall score metrics: Equation 3.11 is used to construct metrics for recall scores by dividing the amount of sentences shared by the reference and candidate summaries..

$$\text{Recall} = \frac{S_{\text{ref}} \cap S_{\text{cand}}}{S_{\text{ref}}} \tag{3.11}$$

BiLingual Evaluation Understudy (BLEU) The mathematical definition of the BLEU score [97] is:

$$\text{BLEU} = \underbrace{\min\left(1, \exp\left(1 - \frac{\text{reference-length}}{\text{output-length}}\right)\right)}_{\text{brevity penalty}} \underbrace{\left(\prod_{i=1}^{4} \text{precision}_{\ i}\right)^{1/4}}_{\text{n-gram overlap}} \tag{3.12}$$

where,

$$\text{precision}_{\ i} = \frac{\sum_{\text{snteCand-Corpus}} \sum_{i \in \text{ snt}} \min\left(m^{i}_{\text{cand}}, m^{i}_{ref}\right)}{w^{i}_{t} = \sum_{\text{snt' } \in \text{ Cand-Corpus}} \sum_{i' \in \text{ snt'}} m^{i'}_{\text{cand}}} \tag{3.13}$$

Here, the total amount of i-grams included in the proposed summary is w^{i}_{t}. The number of i-grams in the candidate that match the reference summary is m^{i}_{cand} and in the reference summary is m^{i}_{ref}.

The BLEU formulation comprise of the brevity penalty and the n-gram overlap component. Brevity Penalty compensates for the absence of a recall phrase in the BLEU score by penalizing summaries that are excessively short relative to the length of the nearest reference with an exponential decay. N-Gram Overlap is a precision metric that measures the number of unigrams, bigrams, trigrams, and four-grams $(i = 1, \ldots, 4)$ that have a corresponding n-gram in the reference summaries. Unigrams account for the summary's sufficiency, whereas longer n-grams account for its fluency. The counts are capped at the maximum n-gram count in the reference $\left(m_{ref}^{n}\right)$, to prevent overcounting.

F-measure: The F-measure is a combined statistic for recall and accuracy as in Equation 3.14. The F-measure represents the middle ground between accuracy and recall.

$$\text{F-Measure } = \frac{2(\text{ Precision })(\text{ Recall })}{\text{Precision } + \text{ Recall}} \tag{3.14}$$

Recall-Oriented Understudy for Gisting Evaluation (ROUGE): It is the most used metric for assessing autonomously generated summaries [98]. A set of measures and a software programme called ROUGE are used in natural language processing (NLP) to assess automated summarization and machine translation technologies [99]. It contrasts various reference summaries produced by people with computer-generated summaries [100]. The main idea behind ROUGE is to measure how many overlapping units—like overlapped n-grams are there between candidate (or system) summaries and reference summaries [101]. ROUGE has been shown to be useful for evaluating summary quality and to correspond well with human judgements [102]. The following are a few ROUGE variations:

ROUGE-1 (R1): It is the distance in unigrams between the candidate's and the references' summaries. An n-gram recall test called ROUGE-N compares candidate and reference summary n-grams.

ROUGE-L (R-L): The longest common subsequences between the candidate and reference summaries are used to assess them.

ROUGE-S* (R-S*): It assesses how many skip-bigrams there are between a candidate summary and a reference summary.

ROUGE-SU* (R-SU*): It is a modification of ROUGE-S* that counts in skipbigrams and unigrams. Asterisks ('*') indicate how many words should be skipped. For instance, ROUGE-SU4 permits bi-grams to include non-adjacent words that are at most four words apart.

Besides these, METEOR and CIDEr are also popular metric in language domain based on the n-gram matching. METEOR relies on an explicit word-to-word match between the assessed MT output and one or more reference translations. In addition, it is also effective for matching synonyms. Likewise, CIDEr aggregates statistics for n-grams throughout the whole dataset using the TF-IDF measure. Intuitively, terms that appear in every caption are less informational and should thus be given less weight when evaluating similarity.

MoverScore uses the semantic separation between a summary and a reference text using the Word Mover's acting on n-gram embeddings collected from BERT representations.

More details on these evaluation metrics can be accessed at Sai et al. [103].

4 RESULTS

Soccer game summaries are of great interest to a variety of parties, including viewers and broadcasters. While other works take into account diverse formats including text, audio, and video, the broadcasting context places a disproportionately greater emphasis on video summary. The objective of this thesis is to provide comprehensive game summaries for soccer matches. The proposed solution requires as inputs a video of a whole soccer match as well as a list of game highlights in the form of event annotations. In contrast, the output is in text form and provides a general overview of the game, as well as a sufficient review of key moments for each game. The output text in English serves as a summary of the soccer game with the maximum value for the total amount of characters being dynamically defined. The emphasis is on text as a modality. In this section, different results are presented. Section 3.3.2 presents three alternative methods that can generate summaries and can be run using the pipeline. Section 4.1 contains the work related to augmentation of various available datasets to make it suitable for their use in summarization task. Section 4.3 describes experiments with the pipeline to generate summary with end-to-end approach.

4.1 Dataset Curation

In this section, the work with the datasets presented in Section 3.2 is described, with indications of how each of them has been augmented as a contribution.

4.1.1 SoccerNet

Ground truth generation: As the SoccerNet dataset as explained in section 3.2.1 is not targeting the game summarization use case, there is no ground truth available for this task. The news, commentary, lineup information, and match statistics for games from multiple leagues from British Broadcasting Corporation (BBC).com were scrapped. The link for each game on BBC's website was carefully curated and only 278 games were found on BBC's website. A web crawler was used to extract the above mentioned information from the web page. The scope of such scrapped news, although wider and containing non-event conversations as well as information irrelevant to that particular game, can be used for training end-to-end systems. Cleaning methods, as well as relevant news selector modules, can be employed to filter the most relevant sentences of interest. This SoccerNet extension will be provided as an open dataset.

4.1.2 SportsSum

Translation: All captions and summaries (news articles) in the SportsSum dataset explained in section 3.2.3 were translated from Chinese to English using Azure Cognitive Services Translator [104].

Metadata extraction: The original SportsSum dataset includes captions and ground truth summaries for each soccer game, but not event metadata. Therefore the few-shot learning capabilities of GPT-3 was exploited to fine-tune the Text-Davinci-001 model with just a few examples to directly output metadata containing the game events in JavaScript Object Notation (JSON) format. The translated game commentaries of 2278 examples were converted to metadata using the fine-tuned GPT-3 model. These examples have commentary, extracted event metadata in each of the commentary lines, and news, all in English. The remaining 3150 examples have translated commentary and news only. The SportsSum extension will be privided as an open dataset.

4.1.3 K-SportsSum

Translation: The detailed captions and game summaries for the 7854 samples in the dataset as mentioned in section 3.2.4 were translated from Chinese to English using Google's Cloud Translation. The captions are in plain English sentences with corresponding timestamps as in the original dataset. Since the metadata extraction has not been performed, the temporal event information is not available. Additionally, 523 sports team information items have also been translated into English. This K-SportsSum extension will be provided as an open dataset.

4.2 Pipeline Implementation

The thesis requires no specific hardware requirements since all the implementations are based on software, and cloud based services are used. STT for automated game commentary extraction utilizes existing solutions, for example, cloud services like IBM Watson. STT can be leveraged with such solutions through API calls through appropriate protocol. Due to large video sizes, batch mode is needed and a Job Queue system needs to be maintained to run STT efficiently. Generative language models with Transformers is used for summary text generation. GPT-3 model API access is needed for finetuning as well as inference for metadata extraction.

The summarization model is trained on a GPU-enabled machine. The setup for the training stage is the machine learning workstation with a generic configuration. Since the datasets that are used by the thesis already contain the video stream, and by the virtue of the thesis scope, any specialized cameras or equipment for data collection purposes are not needed. A web-based Graphical User Interface (GUI) is required to track various stages in the pipeline, and interactivity will be implemented for various actions and flows. The web-based GUI will be hosted on a generic server.

4.3 Experiments

Experiment were conducted with the end-to-end summarization methods described in Section 3.3 using different datasets from Section 3.2, and the ROUGE metric as described in Section 3.4.2. The 3 alternative methods described in 3.3.2 have been executed, and the sample outputs from each of those methods has been collected for evaluation. As ground truth, the game summaries scrapped from BBC.com for SoccerNet, and the English translations generated from the game summaries in Chinese for K-SportsSum were used.

- Method 1.1: The structured metadata are processed through naive templates and the output template output is presented as summary.
- Method 1.2: The structured metadata are processed through naive templates. The output from the template engine is filtered by the audio intensity log and the important ones are selected as per the length budget to be presented as summary.
- Method 1.3: The structured metadata are processed through naive templates. The audio in the video is denoised to obtain human commentary voice only and audio intensity log of the filtered commentator voice is used to select important events.
- Method 2.1: STT is performed on the raw audio obtained form the video and presented as summary.
- Method 2.2: STT is performed on the denoised audio to obtain human commentary voice only and the STT output is presented as summary.
- Method 2.3: STT is performed on the raw audio obtained form the video. The output from STT is filtered by the audio intensity when the and the important ones

are selected as per the length budget.

- Method 2.4: STT is performed on the denoised audio to obtain human commentary voice only. The output from STT is filtered by the audio intensity when the and the important ones are selected as per the length budget.

Sample output summaries form key methods can be found in the Appendix A.6.

4.3.1 Module 2: Noise Suppression on Extracted Audio

As the system's input, the video contains audio data that can be easily extracted. But, such audio inherently contains noise created by the audience and is problematic when it is to be processed by Speech to Text systems to extract commentary. Consequently, noise removal needs to be performed on the audio. Spleeter by Deezer as explained in 3.3.4 is leveraged to split human voices form the audio extracted from videos.

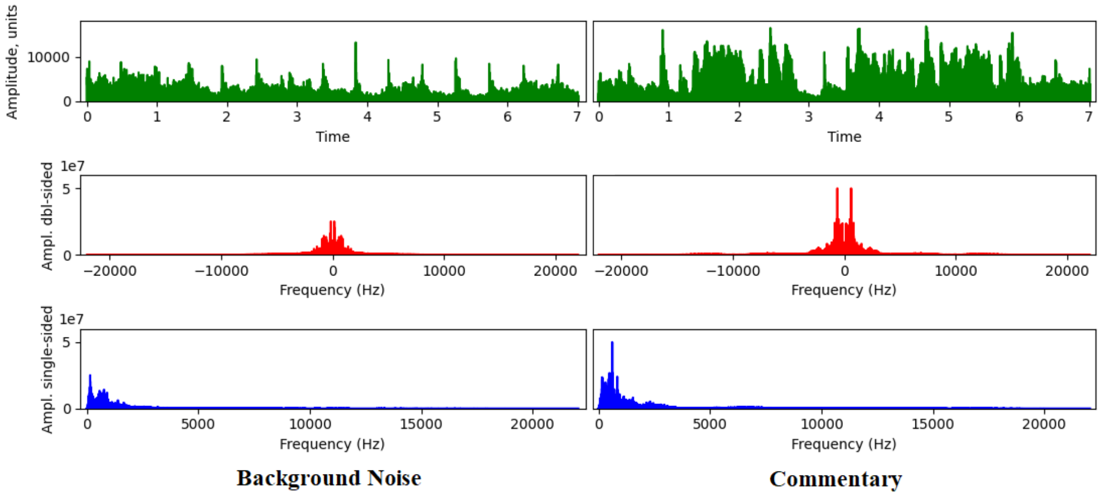


Figure 4.1: Visualization of Extracted Audio on Temporal and Frequency Domain

Figure 4.1 shows the contents on the extracted sound on both temporal and frequency domain. The left part represents the audio where the background noise is only audible and left part represents the audio part where there is live commentary going on. The difference in both the temporal waveform and frequency spectrum on two different cases is clearly visible and this also shows a green signal to the possibilities of separating the noise signal from the commentary to feed this to STT system. Thus, in the final results, the audio feed will be cleaned through noise suppression mechanism and the clean audio will be available.

4.3.2 Module 7: Transformer Model

For method 3.1, a Longformer model trained for multi-document summarization tasks was finetuned for 10 epochs with the maximum output length set to 1024 over 6 NVIDIA V100 GPUs with a per-device batch size of 6. The training took 40 minutes to complete. The model architecture is shown in Figure A.5.

For the experiments, all the translated split-sets in the translated K-SportsSum dataset were combined and filtered such that the character length in the corresponding summary (news article) was less than 2500. Such 7839 samples were again split 80%-20% for training and evaluation. The training loss is shown in Figure 4.2. For the trained model, the attention in multiple heads and layers had been probed. Some sample results are shown in Figure A.2, A.3 and A.4.

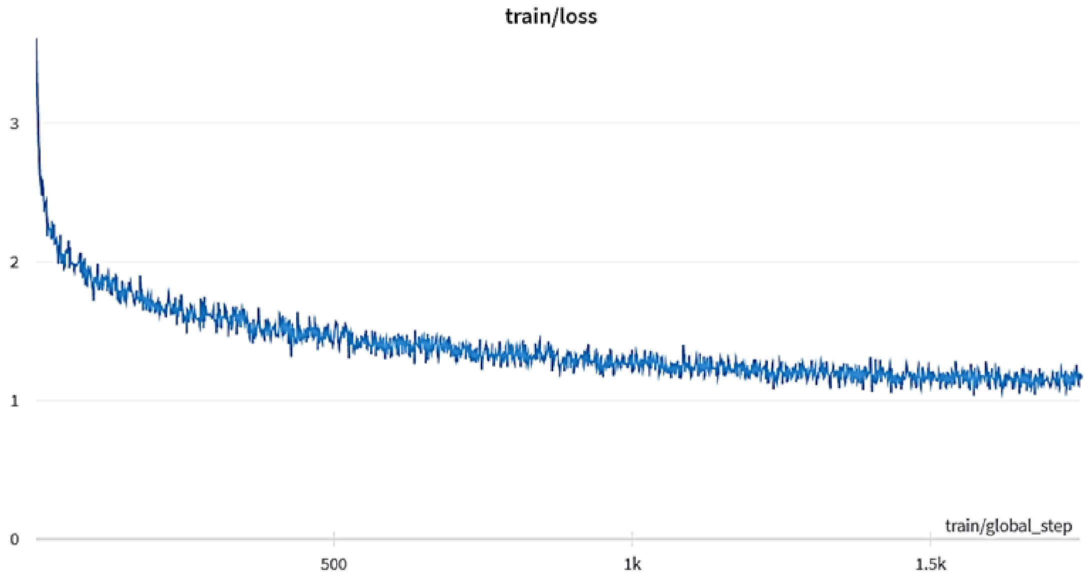
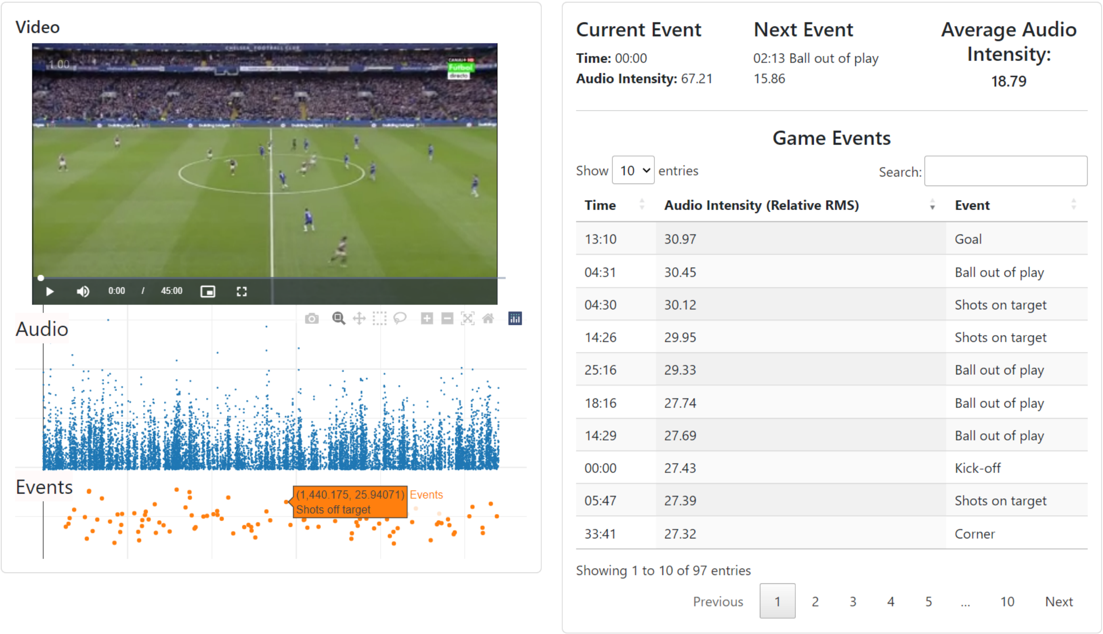


Figure 4.2: Training Loss While Fine-tuning Longformer.

4.3.3 Module 8: Priority Filtering and Audio Dashboard



Current Event
Time: 00:00
Audio Intensity: 67.21

Next Event
02:13 Ball out of play
15.86

Average Audio Intensity:
18.79

Game Events

Show 10 entries Search:

Time	Audio Intensity (Relative RMS)	Event
13:10	30.97	Goal
04:31	30.45	Ball out of play
04:30	30.12	Shots on target
14:26	29.95	Shots on target
25:16	29.33	Ball out of play
18:16	27.74	Ball out of play
14:29	27.69	Ball out of play
00:00	27.43	Kick-off
05:47	27.39	Shots on target
33:41	27.32	Corner

Showing 1 to 10 of 97 entries

Previous 1 2 3 4 5 ... 10 Next

Figure 4.3: Overview of the Audio Dashboard

Priority Filtering module helps to select the key event in the game according to the average RMS intensity of the audio around that event. A frame of seven second is evaluated around each event such that the audio level from before the two seconds before the event as well the the level after five seconds from the event is incorporated for the average RMS calculation. The events are sorted by average RMS intensity and top-N events are selected as the length budget. Alternate strategy like using a threshold level also be adopted to filter the events.

An easy-to-use dashboard is also implemented for visually understanding the correlation between audio intensity levels and important events in the game. The dashboard has interactive charts and tables which helps to navigate through the game events and explore audio intensities along with the events around the time-frame the events had occurred.

4.3.4 Audio Intensity Analysis

In this section, results of audio intensity analysis are presented using the proposed methodology outlined in Sections 3.3.5. The purpose of these specific tests was to test the hypothesis that there is a substantial difference between various event types in terms of the overall intensity of the game audio around the event (including all artefacts such

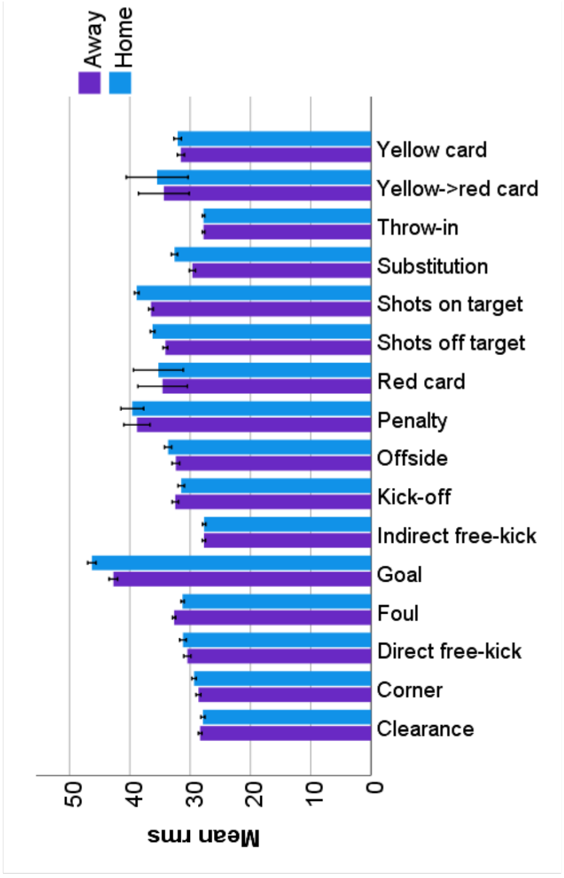


Figure 4.4: Mean Audio RMS Between Different Events Related to Home and Away Team.

as audience cheer, commentator voice, etc.). If yes, what is the magnitude of the total audio intensity for each kind of incident, in order of magnitude? Additionally, the total audio intensity linked to occurrences of the same type may differ depending on whether the event is tied to the home team or the away side. There may also be a change in the overall auditory intensity connected with occurrences depending on whether they occurred in the first half or second half of the game. In terms of frequency composition, wavelet analysis would be used to compare the audio associated with various sorts of occurrences.

Root-mean-square(RMS) Analysis

Table 4.1: Tests of Between-Subjects Effects for Dependent Variable: Audio RMS

Source	df	F	Sig.
Event	15	666.529	0.000
Host	1	6.614	0.010
HalfTime	1	0.004	0.949
Event *Host	15	17.506	0.000
Event * HalfTime	15	2.514	.001
Host * HalfTime	1	0.249	.617

A generalized linear model (GLM) was conducted to compare the RMS in different events and home vs away matches. The results are reported in Table 4.1, where $Event$ indicates 16 different types of events in the dataset, $Host$ indicates whether the event was from a home or away team, and $HalfTime$ indicates whether the event occurred in the first or second half. In the Table 4.1,"df" indicates the degrees of freedom for each variable (number of levels in the variable minus 1). The test statistic from the F-test is called "F" value (the mean square of the variable divided by the mean square of each parameter). The p-value of the F statistic, "Sig.", which represents statistically significant interaction, indicates how probable it is that the F-value computed from the F-test would have happened if the null hypothesis of no difference were true. The results indicate that there is a significant main effect of the event on the RMS, and a significant main effect on home/away matches. This means that, as expected, the generated noise of the crowd was significantly different across various events and if it happened for the home or the away team. There was no significant main effect found for half, meaning that the level of noise from the crowd stayed the same on different half times. However, there is a significant interaction between half and type of event, meaning that there some events create a different noise across two different half times. In addition, a significant interaction found between the hosting and the event, meaning that depending on whether it's home or away, events have a different noise, for example, goals for the home team generated a significantly higher RMS than away matches as shown in 4.4.

Figure 4.5 shows the distribution of audio RMS of different types of events. The higher RMS value represents the higher excitement in the game. The order of importance of

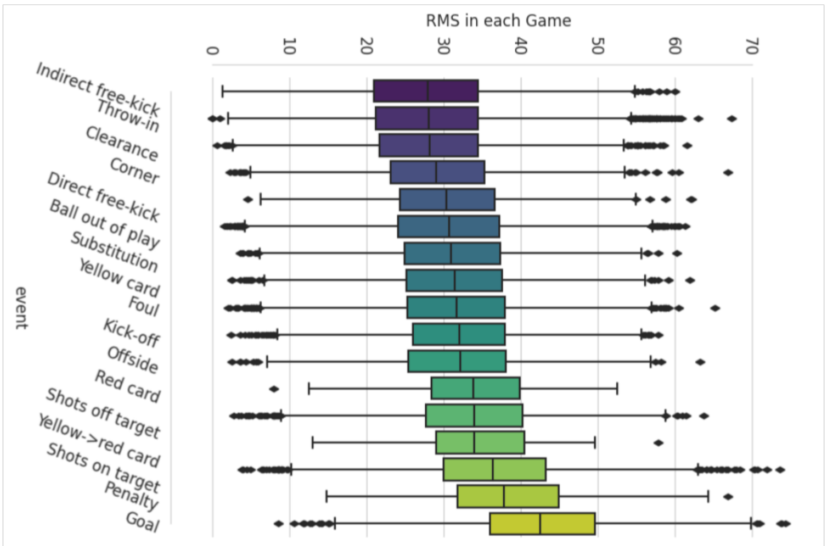


Figure 4.5: The Distribution of Audio RMS of Different Types of Events.

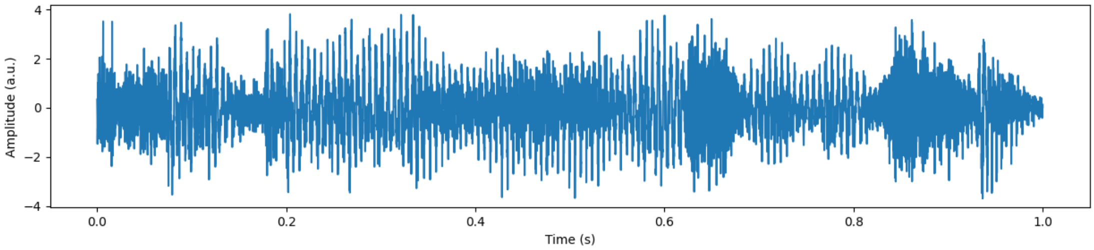


Figure 4.6: 1-Second Audio Time Series After a Goal Event

events across the dataset can be seen in Figure 4.5. It was found that the goal event has the highest RMS intensity followed by the penalty event.

Wavelet Analysis

Using this method, it is explored if there is a difference between the audio associated with different types of events, in terms of frequency composition.

Figure 4.6 shows a time-series of 1-second audio sample from a goal event form the dataset. Figure 4.7 shows the normalized wavelet power spectrum for the goal event with an audio time-series shown in Figure 4.6. The y-axis shows the variation of period, which ranges from the smallest resolvable scale of 2 to 16384, which is around half of the full-time series samples. The hatched area has also been termed the cone of influence, with its conical boundary being the contour line for significance level.

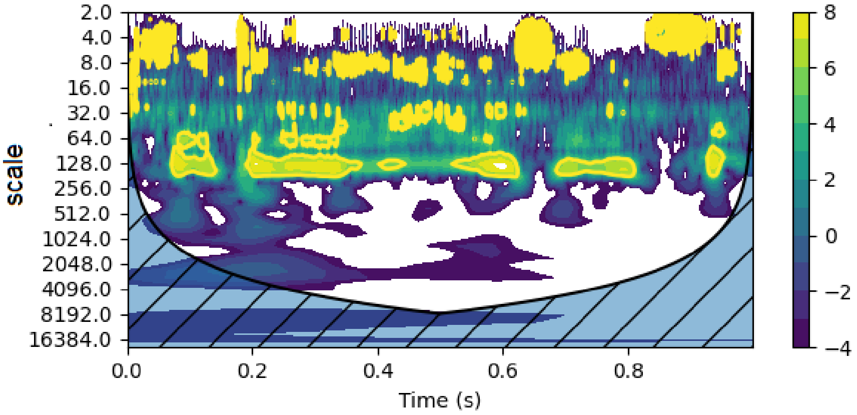


Figure 4.7: Wavelet Power Spectrum

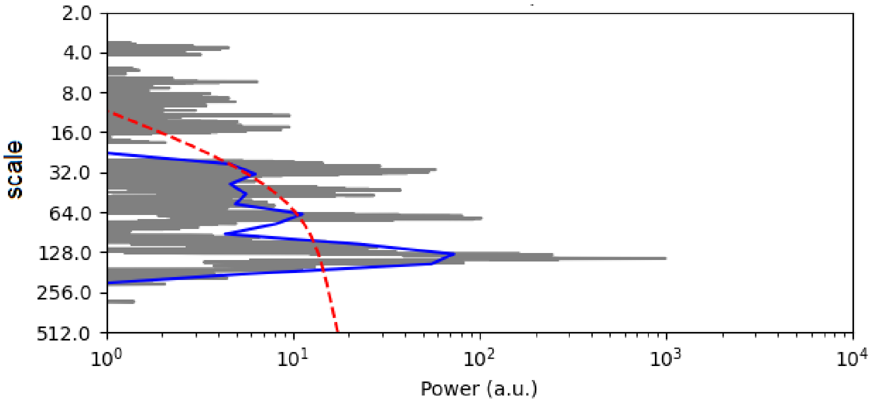


Figure 4.8: Global Wavelet Spectrum.

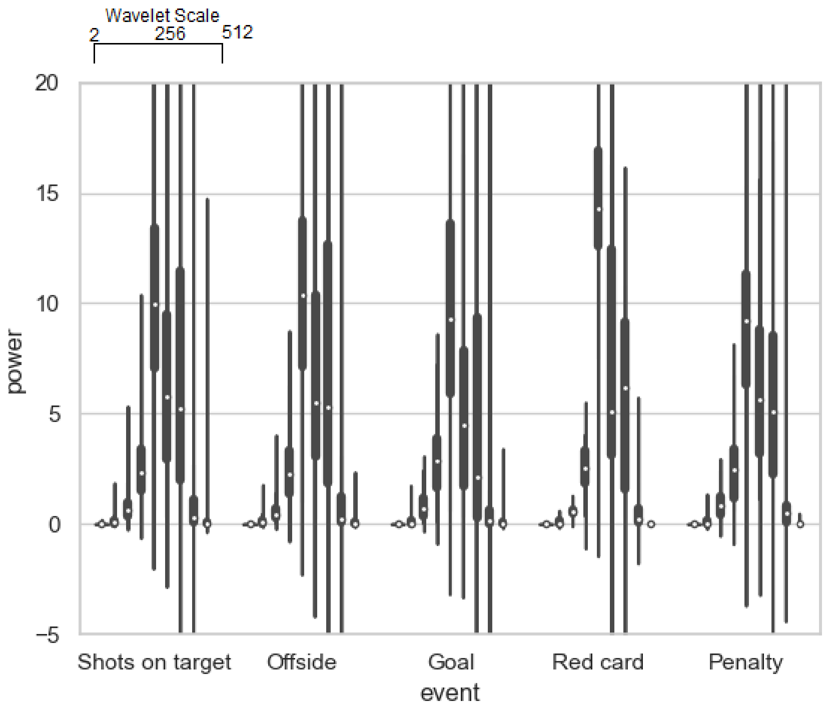


Figure 4.9: Power Distribution Across Wavelet Scales

Figure 4.8 shows the global wavelet spectrum for the goal event with the wavelet power spectrum shown in Figure 4.7. The gray level indicated the Fourier spectrum. The solid line in the figure represents the global wavelet spectrum, with the dotted line reflecting the significance level. The y-axis only shows the variation of period from the smallest resolvable scale of 2 to 512, as the results around the higher order were not much interesting, as seen in Figure 4.7.

Figure 4.9 shows a multiple-box-plot of power values like shown in global wavelet spectrum in Figure 4.7.

Only periods from 2 to 512, which are multiples of 2, are chosen for only five of the different events. A distinct signature across various wavelet scales was observed for different events. For instance, the power of scale 32 was distinctly higher for red-card events when compared to other event types. The power at scales 32 and 128 was seen to be highest in shots, offsides, and goal events.

4.3.5 Length Control

Experiments were carried out to control the length of the summary generated by the Longformer.

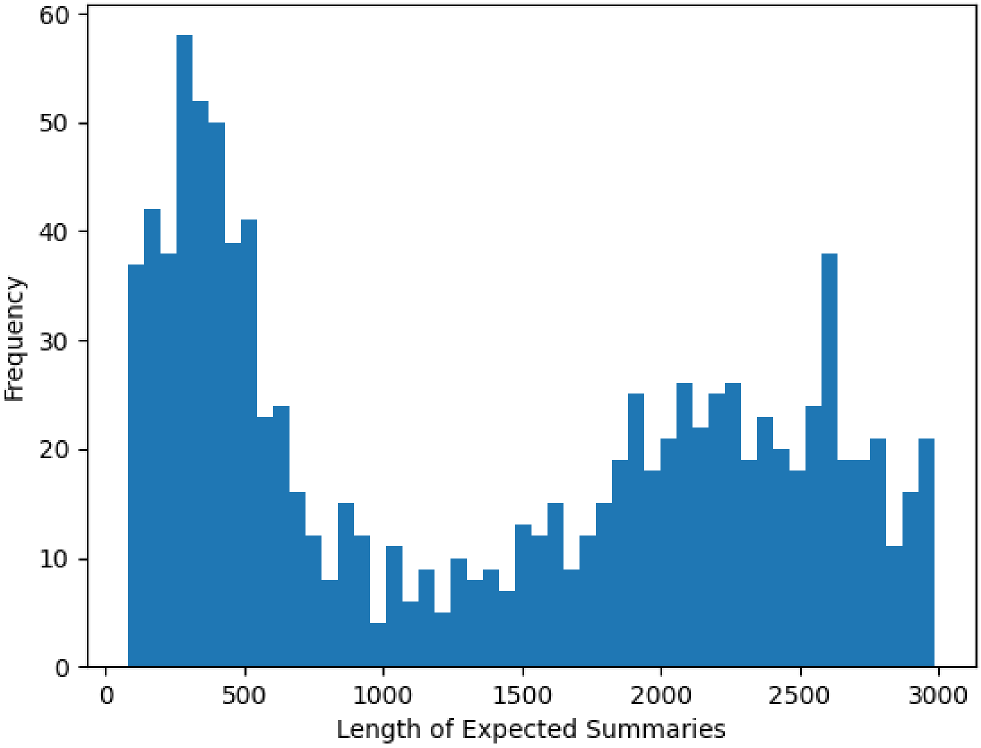


Figure 4.10: Distribution of Length of Summary Sampled From Validation Set

Figure 4.10 represents the distribution of the length of summary in the dataset used to train the Longformer, although this particular histogram is generated only from the 'test' split.

The dataset used to train the transformer contains the translated version of the K-SportsSum dataset, filtered such that the summary is less than 3000 characters. This particular choice is to make sure the expected output from the longformer doesn't exceed the maximum possible length of 1024.

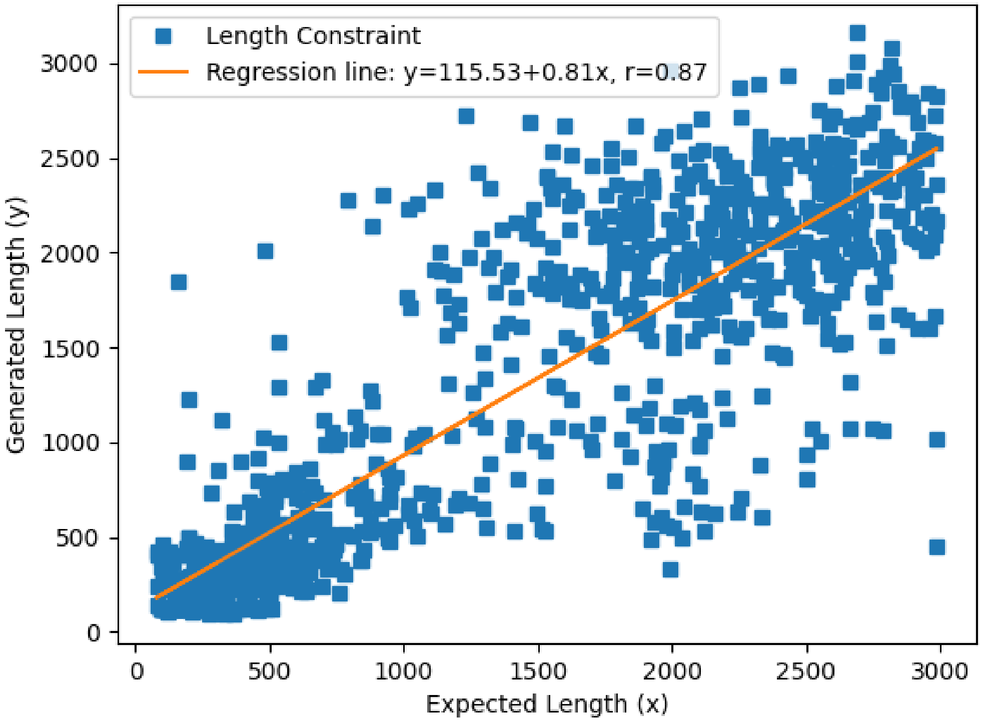


Figure 4.11: Expected Vs. Generated Length of Summary

The information about the expected length constraint was given to the model by tweaking the training dataset. When calculating the length of characters in the summary in the dataset, the information of the length in terms of characters was prepended to the model input. Such augmented data with length information was used to train the model.

Figure 4.11 reflects the capacity of the trained Longformer model to follow the Length Constraint. The x-axis represents the length limit in the number of characters, the model was asked to generate. Likewise, y-axis represents the output length in the number of characters given by the model. A regression analysis was done of the data points obtained by plotting the expected lengths as well as generated lengths. The brown line in the figure represents the line of best fit for the Pearson correlation coefficient of 0.87.

5 DISCUSSION

In this report, the work in progress on the automatic summarization of soccer games in text format has been presented. Through the use of a new end-to-end summarization pipeline, alternative ways for the generation of summaries based on raw game multimedia as well as readily available game metadata and captions, where applicable, utilizing NLP and heuristics, has been explored . A variety of soccer datasets were curated and extended to provided the preliminary findings from the comparative study of different summarization approaches using various input modalities. This has contributed to addressing the outstanding issues in multimodal summarization in a sports context. The open-source software, datasets, and preliminary findings can hopefully be used for future research.

5.1 Summary of Findings and Insights

The proposed method of using a Longformer (Method 3) demonstrates the potential of using transformers for soccer game video summarization. Although they are limited due to the use of fixed-length input, new architectures are suitable for handling longer-term dependencies in the text in both input and output, making them suitable candidates for text summarization. As seen from the results, these approaches can be superior to naive approaches (Methods 1 and 2), since they are ML based methods able to adapt to the input based on their training and configuration, instead of producing static responses.

5.2 Analysis of Results

5.2.1 Objective Analysis

An objective evaluation was performed among the results from various methods. Table 5.1 presents the results in terms of ROUGE-1, ROUGE-2, and ROUGE-L in the real summary outputs from the system.

Table 5.1: Objective Evaluation for Methods 1.1, 1.2, 2.1, 2.4, and 3.1 in Terms of ROUGE-1, ROUGE-2, and ROUGE-L metrics.

Method	Dataset	ROUGE-1	ROUGE-2	ROUGE-L
1.1	SoccerNet	0.08	0.00	0.08
1.2	SoccerNet	0.13	0.01	0.09
2.1	SoccerNet	0.26	0.06	0.10
2.4	SoccerNet	0.29	0.04	0.11
3.1	K-SportsSum	0.52	0.27	0.31

5.2.2 Manual Inspection

Manual inspection of the results was performed at almost every stage during the implementation stage of the thesis. A brief discussion of the observation for methods 1.1, 1.2, 2.1, 2.4, and 3.1 is populated below:

- Method 1.1: As the output is generated form naive template engine, the sentence structures are very similar and boring to read for readers. Since this method contains all the events, the output is complete but very lengthy.
- Method 1.2: Due to the use of priority filters as per audio levels in the game, the summary are shorter and contains key events in the game. But, the output being generated form naive template engine, the sentence structures are still very similar and boring to read for readers.
- Method 2.1: The output is extremely lengthy as it contains STT results on the whole audio of the game. The texts are very difficult to put into context. This might be due to problem in accuracy of STT as well as denoising module.
- Method 2.4: Due to the use of priority filters as per audio levels in the game, summary are shorter and contains texts spoken by commentator when his voice had high RMS energy. The texts are still very difficult to put into context due to problem in accuracy of STT as well as denoising module.
- Method 3.1: The summary is short and contains the key events in the game. The output structure is also very similar to the data the text generation model was trained on. The texts can be related to the real context of the game. However, the information generated by the model cannot always be guaranteed to be true information.

5.2.3 Comparison with the State of the Art

In this section, the work accomplished has been compared in the context of state-of-the-art. The proposed pipeline is the first of its kind in terms of providing an end-to-end automated game summarization functionality requiring no manual intervention in intermedia steps. It is also the first attempt to incorporate audio commentary, metadata, and caption information to provide a comparative analysis. Overall, it can be see that the approach is competitive with, if not better than, existing work in this domain in terms of the objective ROUGE metric.

- **K-SportsSum [81]:** The authors use Selector and Rewriter modules for summarization. They select relevant captions and then rewrite each of them to generate a summary. The proposed work takes the input of whole captions at one once and directly outputs summaries without the need for a selector module. This work has attained a maximum ROUGE score of 48.79).
- **PASS [56]:** This framework heavily depends on detailed metadata and needs pre-specified templates. Essentially, they rely on more detailed metadata as input, and output a naive template-based summarization (akin to Methods 1.2 and 1.3 presented in Section 3.3.2). Presented approach directly generates summaries from readily available captions or the ones genearted with simple templates without the need of structured game and event details.
- **SportsSum [88]:** This work has scrapped news but not cleaned one. In this work, a more advanced version of SportsSum in provided in English translating K-SportsSum.
- **Zhang et al. [105]:** Authors scrapped online sites for events only for their purpose, whereas in this work web scrapping is done for commonly used videos in the open SoccerNet dataset with an additional news component for potential use in summarization.

5.3 Limitations

The sentences in the output generated from the naive template engine are very similar and boring to read for readers. Without the use of priority filters, for both in method family 1 and method family 2, the output contains complete information about the game but is very lengthy. The use of priority filters depends heavily on the audio levels in the game. The audio intensity of an event is extracted from the game video as per the timestamps available in the metadata. Therefore, the effectiveness of this method is directly related to the accuracy and precision of the timestamps available in the meta-data.

Method family 2 heavily depends upon the quality of STT results, which in turn depends upon the quality of the audio available in the video. The quality of the audio available also plays a great role in the performance of the denoiser module.

The output of a trained transformer model for summary generation depends upon the dataset available to train the module. It has been seen that the scope of the content in the news can greatly affect the performance of the model training.

6 FUTURE ENHANCEMENTS

In this section, we note the open challenges in the domain of soccer game summarization, as well as the various limitations and shortcomings of our work.

6.1 Open Challenges

- **Lack of open datasets:** There are no public unified datasets available with all the information like game videos, commentaries, event information, and related game news. Commentaries, event information, and news are available on online websites and portals. The news is mostly of wider scope than the particular game itself. However, for training a summarization algorithm, only relevant sentences about the game are preferred.

- **Heterogeneity in-game metadata:** Although metadata generation is part of commercial broadcast pipelines, open datasets are far from having access to such information. In cases when game metadata is available, it is mostly sparser and heterogeneous among datasets in terms of types of events available, details, etc.

- **Multilingual data** Multiple languages used in game broadcasts, as well as news, can be a problem while dealing with STT as well as summarization systems. STT and translation systems are evolving to handle a wide array of languages, giving hope that future models will have multilingual understanding and generation capabilities accommodating multilingual inputs as well as outputs.

- **Noise removal** Various noises are inherently present in the game audio. Prevailing noise removal mechanisms are not suitable out-of-the-box for soccer game audio. Domain adopted methods to remove noise effectively from the commentary audios of game videos would be beneficial for further downstream tasks like STT.

6.2 Future Work

Tasks that will not be feasible to complete within the given timeframe for this thesis, but are nonetheless interesting as potential future work topics includes:

System Design and Conceptualization

System will be conceptualised and designed and to reflect the pipeline to represent the multiple ways of generating summary from available multi-model input.

Dataset Curation

Dataset exploration and curation will be continued to make richer dataset for the experiments as well as contributions.

Pipeline Improvements

The pipeline would be continually improved with additions and modification of current components and models. GPT can be fine-tuned for metadata extraction from the caption. Longformer can be fine tuning with more data and global attentions.

Performance Evaluation

More metric will be used for performance evaluations like BLUE and some other embedding-based metrics. The focus will be on subjective evaluation as well using real human and the experience will be quantified as possible.

7 CONCLUSION

This work explored the domain of automatic summarization of soccer games in text format. A pipeline for game summarization was proposed to utilize multi-model data as an input to the system. The Longformer model was found to be effective in being used for the summarization of game events. Existing datasets were curated and expanded from a summarization viewpoint. Using RMS and wavelet analysis on audio data, it was also investigated how audio intensity can be utilised for event filtering and prioritization for summarization.

In line with the open challenges listed above, there are a number of directions for potential future work. The plausible direction can be around dataset curation, which also includes the analysis, cleanup, and validation of the translations in multiple datasets. The human augmented or maybe automated ways to clean the news scraped from the internet that has a large scope and contains information irrelevant to the game can be explored. The effectiveness of using a noise-reduction system before doing STT can be studied. The direction around meta-data parsing from existing commentary to make an existing dataset rich in information can be explored. The possibilities of extracting meta-data directly from human commentary from the audio in the game video can also be investigated. The efficiency of using a selector module to select relevant commentaries before feeding them to the transformer model can be inspected, which might be very effective in limiting the input size to the model. Better mechanisms to explicitly control the output of the text generator module can be explored. Subjective user studies, which are more informative about the end-user experience than the objective ROUGE scores, can be conducted to measure the effectiveness of the system.

A APPENDIX

A.1 Terminology

Below is the terminology which has been used throughout this work.

- **Soccer:** Also called association football[1], played in accordance with a codified set of rules known as the Laws of the Game (LOTG)[2] by the International Football Association Board.
- **Broadcast:** A transmit of television.
- **Streaming:** Different than broadcasting, Over-the-Top (OTT) delivery of content.
- **Game/match:** Game is any soccer play including unofficial/amateur meetings, where match refers to official competitions.
- **Football club:** An organization of players, managers, owners or members associated with a specific football team[3].
- **Association club:** Refers to a particular country, and the national team.
- **Event:** (In the context of multimedia content, not to be confused with the entire soccer game.) Also called "highlight". According to the Cambridge Dictionary an event is defined as anything that happens, especially something important or unusual[4]. In this work we refer to important happenings in a soccer match relevant to the game outcome, such as *goal* (when the ball passes the goal line), *card* (when the referee hands a yellow or red card to a player), *substitution* (when one player is substituted off and another player goes on the pitch), and many others including *free kick*, *corner*, *offside*, and *penalty*.
- **Highlight clip:** Video clip displaying a particular event from a soccer game.
- **Highlight reel:** Combination of highlight clips (video-based type of a game summary).

[1] https://en.wikipedia.org/wiki/Names_for_association_football

[2] https://www.fifplay.com/downloads/documents/laws-of-the-game-2021-2022.pdf

[3] https://www.lexico.com/definition/football_club

[4] https://dictionary.cambridge.org/dictionary/english/event

- **Tagging/annotation:** Setting timestamps on events in soccer match, adding metadata, etc.

- **Tagging center:** A typical tagging center in live operation is shown in Figure 1.1.

- **Game summary (text):** A text describing selected events and highlights from a soccer game.

- **Background noise:** The audience cheer and musical instruments played during the games, apart from the voices of commentators.

A.2 Thesis Schedule

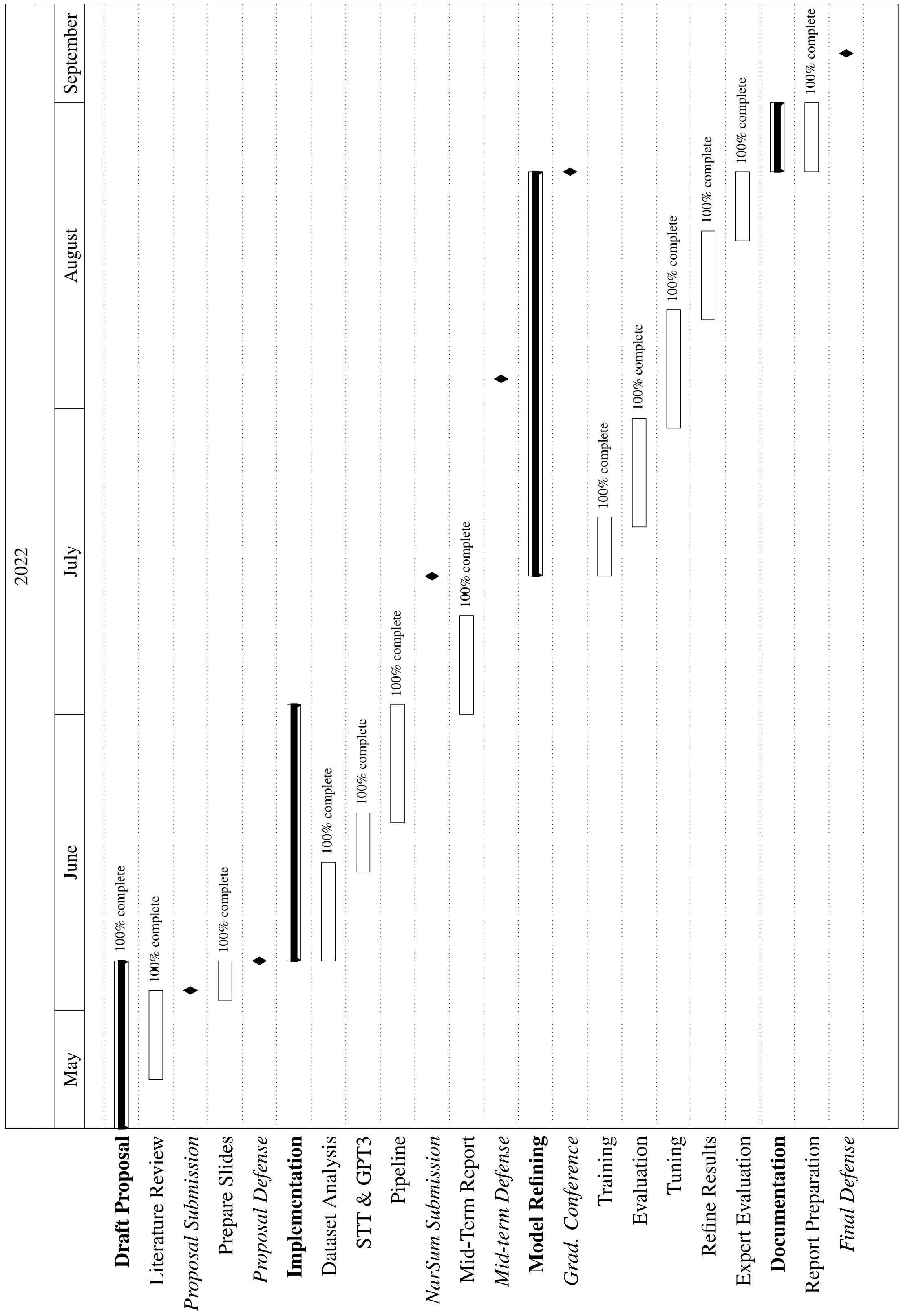


Figure A.1: Gantt Chart Showing the Thesis Timeline

A.3 Literature Review of Base Papers

Literature Review of Base Paper- I

Author(s)/Source: Samah Aloufi, Abdulmotaleb El Saddik	
Title: MMSUM Digital Twins: A Multi-view Multi-modality Summarization Framework for Sporting Events	
Website: `https://doi.org/10.1145/3462777`	
Publication Date: January 2022	**Access Date: May 2022**
Journal: ACM Transactions on Multimedia Computing, Communications, and Applications	**Place:** n.a.
Volume: Volume 18, Issue 1	**Article Number:** Article No.: 5
Author's position/theoretical position: Assistant Professor	
Keywords: Information retrieval; Multimedia information systems; Machine learning;	
Important points, notes, quotations	**Page No.**
1. Textual Summary: selects the top ranked tweets	**11**
2. Visual Summary: select the topN ranked images based on popularity score	**11**
3. a sentiment analysis aspect which tracks the changes in fan sentiment in correlation to event happenings, and, predicting which images will be more popular than others during a new fresh event to be utilized in the summary	**9**
4. a multi-view aspect which describes the event differently based on individual fan perspectives, a sentiment analysis aspect which tracks the changes in fan sentiment in correlation to event happenings, and, predicting which images will be more popular than others during a new fresh event to be utilized in the summary	**11**
5. Found goals, penalties, red cards, and own goals are key events that will affect fan satisfaction with a football game summarization.	**17**
Essential Background Information: Timely shared text posts, images, and videos on social media captures on-the-spot information, and represents people's opinions, sentiments, views, and reactions, which center around specific events can be utilized to generate game summaries.	
Overall argument or hypothesis: Chronological multi-view multi-modal summarization can be done utilizing microblog stream during the playing time of a given match to update the users with the occurrence of significant sub-events such as goals, red cards, or penalties, and how fans react to these sub-event based on their perspectives	
Conclusion: Formulated a sporting-event Digital Twins summarization framework (MMSUM) developed to generate a multi-view multi-modality summary of a football (soccer) match in near real-time with various components required for event stream monitoring and analysis.	
Supporting Reasons	
1. Evaluated summarization approach on real world data collected during the 2018 FIFA World Cup.	**2.** User study to assess the quality and the multi-view aspect of generated summaries.
3. 78% Accuray using Kernel SVM on Images Types Classification	**4.** Algorithm is able to detect most of the goals, own-goals, penalties, and red cards with High F1 score.
5. Automatically generated summary achieved 48% in terms of recall and 32% in term of F-measure using ROUGE-1.	**6.** 77% of the participants who evaluated the final match agreed that the length of the provided summary was adequate
7. 73% of the participant agreed that the presented tweets (images) are related to the match.	**8.** 69% of the participants who evaluated the final match agreed that our summary does convey the same information as the ESPN.com
Strengths of the line of reasoning and supporting evidence: conducted a user study to assess the quality and the multi-view aspect of the generated summaries. The evaluation results show the promising potential of our approach in providing subjective summarizations based on fans' varying points of view.	
Flaws in the argument and gaps or other weaknesses in the argument and supporting evidence: Multiple as well as colloquial languages are found in social media but there is no way to handle/utilize such data streams. No provision to handle unplanned events.	

Literature Review of Base Paper- II

Author(s)/Source: Lorenzo Gatti, Chris van der Lee, Mariët Theune

Title: Template-based multilingual football reports generation using Wikidata as a knowledge base
PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences

Website: `http://dx.doi.org/10.18653/v1/W18-6523`
`http://dx.doi.org/10.18653/v1/W17-3513` **(Original PASS system)**

Publication Date: November, 2018 | **Access Date: May, 2022**

Journal: The 11th International Natural Language Generation Conference | **Place:** The Netherlands

Volume: 11th | **Article Number:** n.a.

Author's position/theoretical position: PostDoc at University of Twente

Keywords: knowledge base; summary template; machine translation

Important points, notes, quotations	**Page No.**
1. extended PASS to produce English texts, exploiting machine translation	**183**
2. Wikidata as a large-scale source of multilingual knowledge	**184**
3. Multilingual NLG research normally needs additional grammar related work	**183**
4. can generate tailored emotional language	**184**
5. manually corrected translated templates with aid of tools	**184**

Essential Background Information: They extended existing Football reports generation system called PASS that generated Dutch text and relied on a limited hand crafted knowledge base, to produce English texts, exploiting machine translation and WikiData.

Overall argument or hypothesis: Machine Translation can be used to translate templates from one languages to another. Knowledge present in Wikidata can supplement hand-crafted local knowledge base.

Conclusion: Wikidata turned out to be an useful resource, thanks to its extensive coverage - both in terms of knowledge and languages present - and its dynamic nature, and that it should be considered when developing multilingual NLG systems. But! better templates could be produced by repeating the same methodology used for creating the original PASS templates, i.e. starting from a corpus of English reports and manually annotating some sentences, replacing the entities therein contained with placeholders. MT was not so perfect for them back then.

Supporting Reasons

1. A human-based evaluation to measure the text quality of PASS. 20 game reports (2 per soccer match) were evaluated by each of participants.

2. Evaluation showed (in another paper) that readers were clearly able to distinguish the team for which a report was tailored in 91% of all cases.

3. The acceptable levels of clarity and fluency for the reports, while the correctness of the information given was higher than in human-written reports.

4. A chi-square test also showed a significant correlation between the intended and perceived tailoring towards fans of the clubs $\left(\chi^2(1) = 233.33, p < .001\right)$.

5. Participants were overall positive in regards to the clarity and fluency of the reports. The average scores of clarity $(M = 5.64, SD = 0.88)$ and fluency $(M = 5.36, SD = 0.79)$ were well above the neutral score of 4 .

Strengths of the line of reasoning and supporting evidence: Human-based evaluation to measure the text quality produces by PASS. No Objective measure to compare with previous system as the system is itself noble in its domain.

Flaws in the argument and gaps or other weaknesses in the argument and supporting evidence:
Automatic translations has problems with idiomatic expressions.
Lexical variations was not applied after generating the sentence from a template. There is an opportunity to use variations in the generated sentences.
This evaluative content is currently not backed up by objective data, but can be seen as the subjective view in favor of one side

Literature Review of Base Paper- III

<table>
<tr><td colspan="2"><b>Author(s)/Source:</b> Wang, Jiaan and <b>Li, Zhixu</b> and Zhang, Tingyi and Zheng, Duo and Qu, Jianfeng and Liu, An and Zhao, Lei and Chen, Zhigang</td></tr>
<tr><td colspan="2"><b>Title:</b> Knowledge Enhanced Sports Game Summarization</td></tr>
<tr><td colspan="2"><b>Website:</b> https://dl.acm.org/doi/10.1145/3488560.3498405</td></tr>
<tr><td><b>Publication Date: February 2022</b></td><td><b>Access Date: May, 2022</b></td></tr>
<tr><td><b>Journal:</b> Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining</td><td><b>Place:</b> New York, NY, USA</td></tr>
<tr><td><b>Volume:</b> WSDM ’22</td><td><b>Article Number:</b> Pages 1045–1053</td></tr>
<tr><td colspan="2"><b>Author’s position/theoretical position:</b> Postdoctoral Fellow, KAUST, SA, etc</td></tr>
<tr><td colspan="2"><b>Keywords:</b> datasets, sports game summarization, text summarization</td></tr>
<tr><td colspan="2"><b><u>Important points, notes, quotations</u></b> <b><u>Page No.</u></b>
1. Introduce K-SportsSum, a new dataset from massive games <b>1</b>
2. Manual cleaning process in commentary-news pairs to improve the quality, <b>1</b>
3. a large-scale knowledge corpus that contains the information of 523 sports teams and 14,724 sports player <b>1</b>
4. a knowledge-enhanced summarizer that utilizes both live commentaries and the knowledge to generate sports new <b>1</b></td></tr>
<tr><td colspan="2"><b>Essential Background Information:</b> Attracted attention from both the research communities and industries to generate report corresponding news articles after games. Generated sports news should record the core events of a game that could help people efficiently catch up to games.</td></tr>
<tr><td colspan="2"><b>Overall argument or hypothesis:</b> Proposed a knowledge-enhanced summarizer that first selects key commentary sentences, and then considers the information of the knowledge corpus during rewriting each selected sentence to a news sentence so as to form final news.</td></tr>
<tr><td colspan="2"><b>Conclusion:</b> The experimental results on K-SportsSum and SportsSum datasets show that the model achieves new state-of-the-art performances. Wualitative analysis and human study was done to verify that the model generates more informative sports news</td></tr>
<tr><td colspan="2"><b>Supporting Reasons</b></td></tr>
<tr><td><b>1.</b> the model outperforms the baselines on both K-SportsSum and SportsSum datasets in terms of ROUGE scores</td><td><b>2.</b> The model made use of an additional corpus to alleviate the knowledge gap.</td></tr>
<tr><td><b>3.</b> The model used a pre-trained language model (mT5) which made use of the knowledge embedding to implicitly generate informative sports news.</td><td><b>4.</b> The model had a better performance on generating sports news in terms of informativeness, fluency and overall quality.</td></tr>
<tr><td><b>5.</b> The model was able to generate correct descriptions of the sports events.</td><td><b>6.</b> The model was able to take the different text styles into account.</td></tr>
<tr><td colspan="2"><b>Strengths of the line of reasoning and supporting evidence:</b> The paper proposes a knowledge enhanced summarizer to harness external knowledge for generating more informative sports news. They conducted extensive experiments to verify the effectiveness of the proposed method on two datasets compared with current state-of-the-art baselines via quantitative analysis, qualitative analysis and human study.</td></tr>
<tr><td colspan="2"><b>Flaws in the argument and gaps or other weaknesses in the argument and supporting evidence:</b> They did the experiments in Chinese corpus. The result in English language can be different. The model can behave differently on generating short texts. There is no explicit mechanism for length control in the model proposed.</td></tr>
</table>

Literature Review of Base Paper- IV

Author(s)/Source: Cise Midoglu, Steven A. Hicks, Vajira Thambawita, Tomas Kupka, Pål Halvorsen

Title: MMSys'22 Grand Challenge on AI-based Video Production for Soccer

Website: `https://arxiv.org/abs/2202.01031`
`https://mmsys2022.ie/authors/grand-challenge`

Publication Date: June 2022	**Access Date: May, 2022**
Journal: 13th ACM Multimedia Systems Conference (MMSys'22)	**Place:** Athlone, Ireland.
Volume: 13th	**Article Number:** n.a.

Author's position/theoretical position: Simula Team

Keywords: machine learning, soccer, video clipping, thumbnail selection, video summarization

Important points, notes, quotations **Page No.**

1. Existing works consider different modalities such as video, audio, and text, but a relatively larger emphasis is put on video summaries in the broadcasting context. **1**
2. Video and metadata from the Norwegian Eliteserien are used, and submissions will be evaluated both subjectively and objectively **5**
3. a potential for AI-supported clipping of sports videos, especially in terms of extracting temporal information. **2**
4. automatic eaningful and attractive thumbnails thumbnail selection system exploits two important characteristics: high relevance to video content, and superior visual aesthetic quality. **3**

Essential Background Information: Automated solutions might enable leagues to be broadcasted and/or streamed with less funding, at a cheaper price to fans.

Overall argument or hypothesis: Soccer game summaries are of tremendous interest for multiple stakeholders including broadcasters and fans but a relatively larger emphasis is put on video summaries in the broadcasting context

Conclusion: Recent methods can be integrated to create a pipeline for video summarization system and it has a huge potential to be used in the practical context.

Supporting Reasons

1. demonstrates the applicability of ML, and more specifically adversarial and reinforcement learning

2. However, a lot of work remains to be done for the implementation of automated algorithms within the soccer domain

3. audio may be an important modality for finding good clipping points and also may be to find important events

4. a potential for AI-supported clipping of sports videos, especially in terms of extracting temporal information

5. presented results are still limited, and most works do not directly address the actual event clipping operation

6. computing should be possible to conduct with very lowvlatency, as the production of highlight clips needs to be undertakenvin real-time for practical applications

7. traditional solutions in the soccer domain rely on the manual or static selection of thumbnails to describe highlight clips, which display important events such as goals and car

8. Manual clipping result in the selection of sub-optimal video frames as snapshots, which degrades the overall quality of the clip as perceived by the viewers, and consequently decreases viewership. Additionally, manual processes are expensive and time consuming

Strengths of the line of reasoning and supporting evidence: The paper clearly explains the research gap in the soccer video domain and clearly indicates that text summarizing of soccer game videos with new tools is relatively unexplored topic.

Flaws in the argument and gaps or other weaknesses in the argument and supporting evidence: Haven't explained the works in other domain rather than soccer. For example, there have been researches om data-to-text models.

A.4 Annotation Structure of SoccerNet Metadata

Listing A.1: Annotation Structure for Sample Event in Action Spotting Task of SoccerNet-V2.
(*) Field optional.

```
{
"gameTime"   : "<half Number> - <time from start of the half>",
"label"      : "<action type>",
"position"   : "<time in ms from beginning of game>",
"team"       : "<home/away/not applicable>",
"visibility": "<visible/not shown>"

}
```

Listing A.2: Data Structure for Action Spotting Task of SoccerNet-V2.

```
# Start
{
    "UrlLocal"    : "<path>",
    "UrlYoutube"( * ) : "<link>",
    "annotations" :[{
        ...

    }],
# End
    "gameAwayTeam": "<team name>",
    "gameDate": <timestamp>,
    "gameHomeTeam": "<team name>",
    "gameScore": "<int> - <int>"
}
```

A.5 Annotation Structure of HOST Metadata

Listing A.3: Annotation Structure for Sample Card Event.

```
{ '<timestamp>',
    '{ "team":
            {
                "id"     : <team-id>,
                "type"   : "team",
                "value"  : "<team-name>"
            },
        "action": "<yellow/red> card",
        "player":
            {
                "id"     : <player-id>,
                "type"   : "player",
                "value"  : "<player-name>"
            }
    }'
}
```

Listing A.4: Annotation Structure for Start and End Timestamps.

```
# Start
{ '<timestamp>',
    '{ "phase":
            {
                "type"   : "phase",
                "value"  : "<1st/2nd> half"
            },
        "action": "start phase"
    }'

}
# End
{ '<timestamp>',
    '{
        "action": "end of game",
    }'
}
```

Listing A.5: Annotation Structure for Sample Goal Event.
(*) Field optional.

```
{ '<timestamp>',
    '{ "team":
            {
                "id"    : <team-id>,
                "type"  : "team",
                "value" : "<team-name>"
            },
        "action": "goal",
        "scorer":
            {
                "id"    : <player-id>,
                "type"  : "player",
                "value" : "<player-name>"
            },
        "assist by":
            {
                "id"    : <player-id>,
                "type"  : "player",
                "value" : "<player-name>"
            },
        "shot type":
            {
                "type"  : "goal shot type",
                "value" : "<shot-type>"
            }
        "after set piece"(*):
            {
                "type": "set piece",
                "value": "penalty"
            }
    }'
}
```

A.6 Sample Summaries Generated Through the Pipeline

Pipeline Output from Method 1.2

summary from game metadata using naive template

Ball out of play happened in 3rd minutes. Corner from Swansea in 4th minutes. Shots on target from Swansea in 4th minutes. Ball out of play happened in 9th minutes. Ball out of play happened in 19th minutes. Clearance from Manchester United in 21st minutes. Goal from Manchester United in 27th minutes. Goal from Swansea in 29th minutes. Kick-off from Manchester United in 30th minutes. Ball out of play happened in 31st minutes. Ball out of play happened in 32nd minutes. Throw-in from Swansea in 34th minutes. Ball out of play happened in 34th minutes. Foul from Manchester United in 37th minutes. Offside from Swansea in 41st minutes. Yellow card from Manchester United in 44th minutes. Substitution from Manchester United in 0th minutes. Offside from Swansea in 2nd minutes. Yellow card from Manchester United in 5th minutes. Corner from Manchester United in 9th minutes. Substitution from Manchester United in 25th minutes. Ball out of play happened in 30th minutes. Offside from Manchester United in 35th minutes. Goal from Swansea in 54th minutes. Substitution from Swansea in 57th minutes. Substitution from Manchester United in 66th minutes. Ball out of play happened in 67th minutes. Yellow card from Manchester United in 77th minutes. Shots on target from Swansea in 80th minutes. Yellow card from Swansea in 84th minutes. Yellow card from Manchester United in 85th minutes. Foul from Swansea in 90th minutes. Yellow card from Swansea in 90th minutes. Substitution from Swansea in 95th minutes.

Pipeline Output from Method 2.4

summary from game audio using naive STT transcription

In the area he loves today . And he is happening since this month yes united states has changed the since looking for a little bit of revenge for diamond shadow play before to remember any second support to strike you so talking auctions to full structural . Not so cheap three gigs . The smart finish . So we have to say . One minute’s second this between . Last season a really successful time as well and wanted to take it but he feels that this is his home and gary monkey for lengthy discussions . These try to find party mean area . If so just caught the wrong side of his mom . S- delivery is trust . Quite readjustment is needed . Yesterday we want the feet in the last nineteen ames . Situation behind the solution . The past has been plenty of sign for manchester market they’ve never panicked in these situations . How it’s john joselbe . Immediately apologises i think he tried to actually . On the banks that he may not get the goals . The swansea city now push everybody back all hands to the pump . Just to stop us to quite rise the correctly .

Pipeline Output from Method 3.1

summary from game commentary using transformer model

After the opening Jeonbuk Hyundai took the lead with its first shot. In the 12th minute Eninho made a cross from the right and Han Kyo-won headed the ball into the lower right corner from the front of the penalty area. In halftime stoppage time Jeonsang Hyundai took a right corner kick from the left and Eninho's left-footed shot was blocked by the opponent's defender. In half-time stoppage Je Jeonba Hyundai took another right corner and Jeonbaruk Hyundai's left foot shot from outside the opponent's penalty area was saved by the opposing goalkeeper. After the first half of the game was over the two sides changed sides and fought again and the situation remained the same. After halftime the game remained in a stalemate with the two teams fighting fiercely and neither was able to create any scoring opportunities. The two sides finally shook hands and fought successively with a 1-1 draw at home.

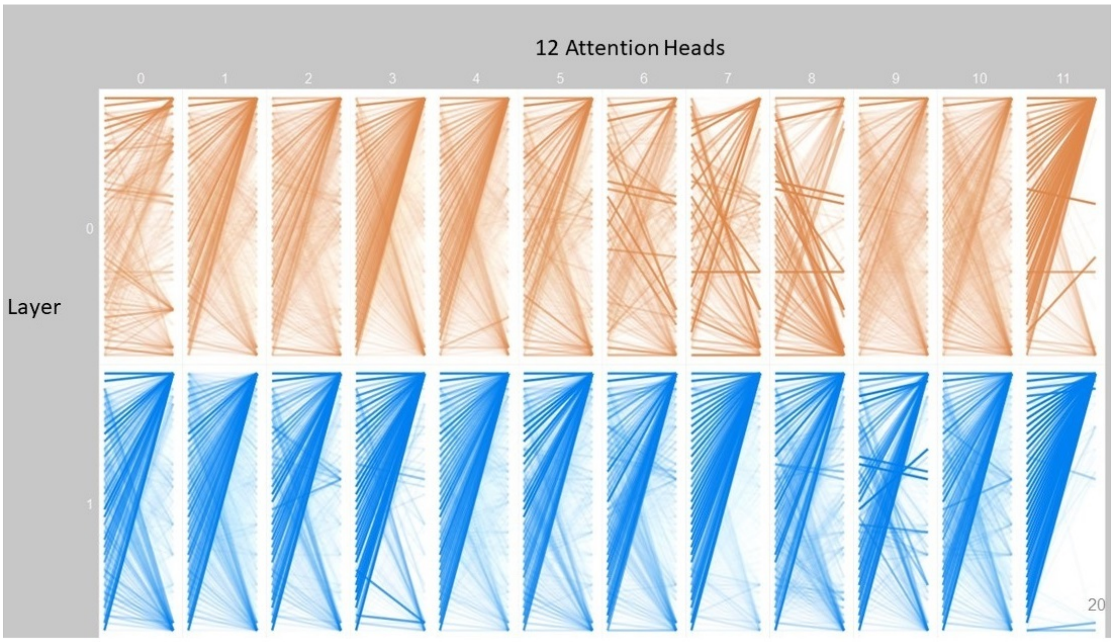


Figure A.2: Exploring Attention in Trained Longformer-1

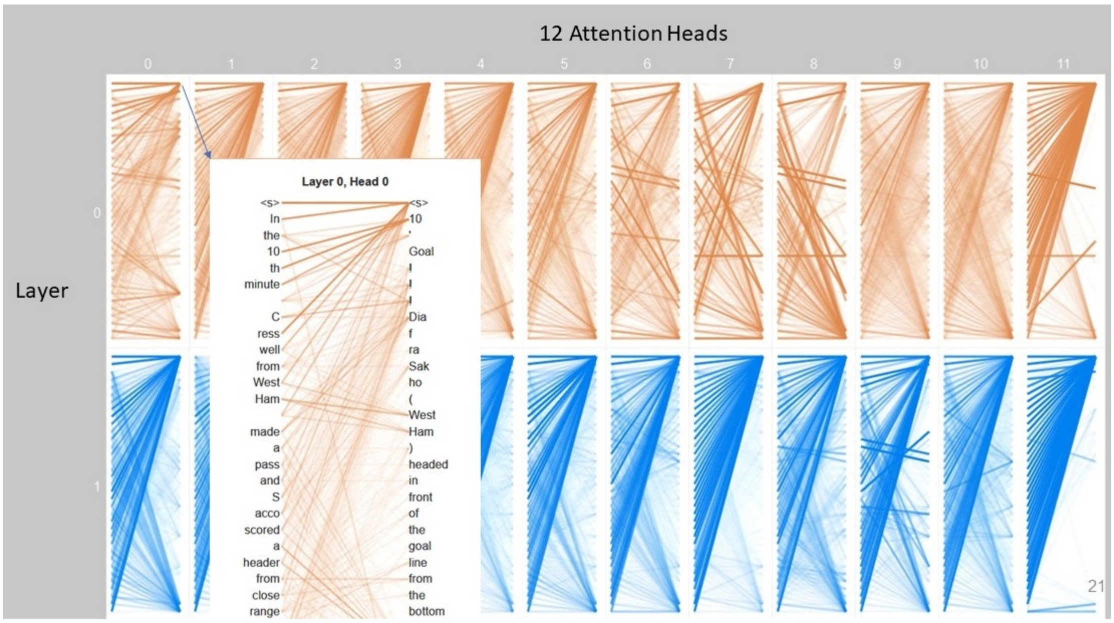


Figure A.3: Exploring Attention in Trained Longformer-2

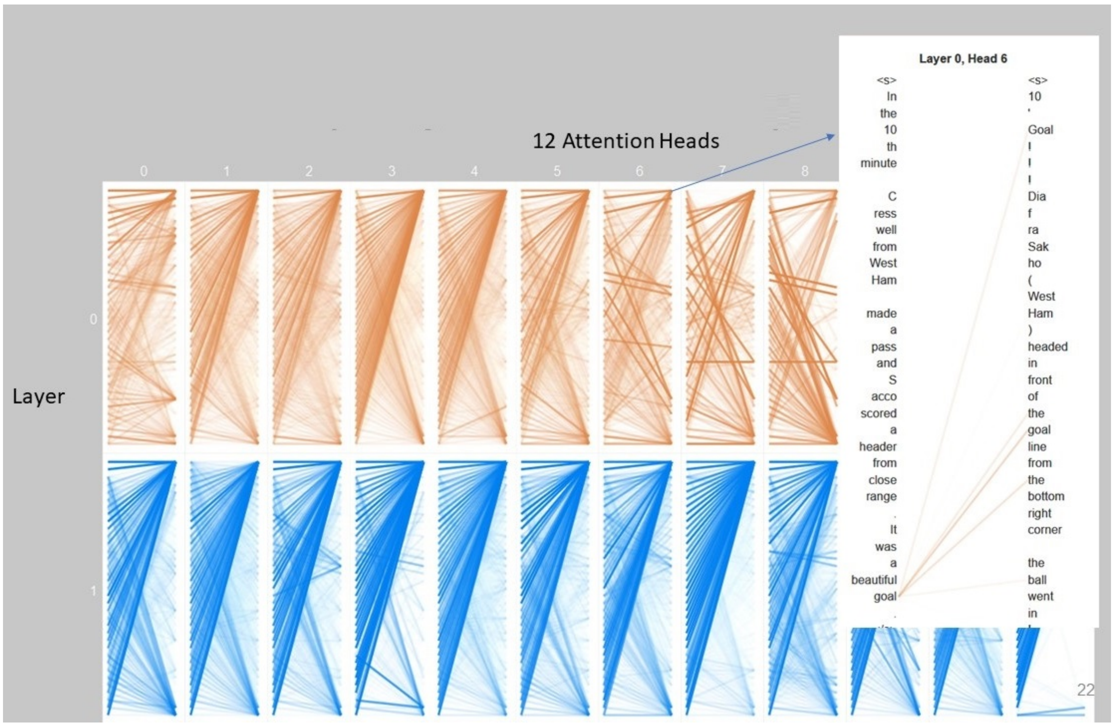
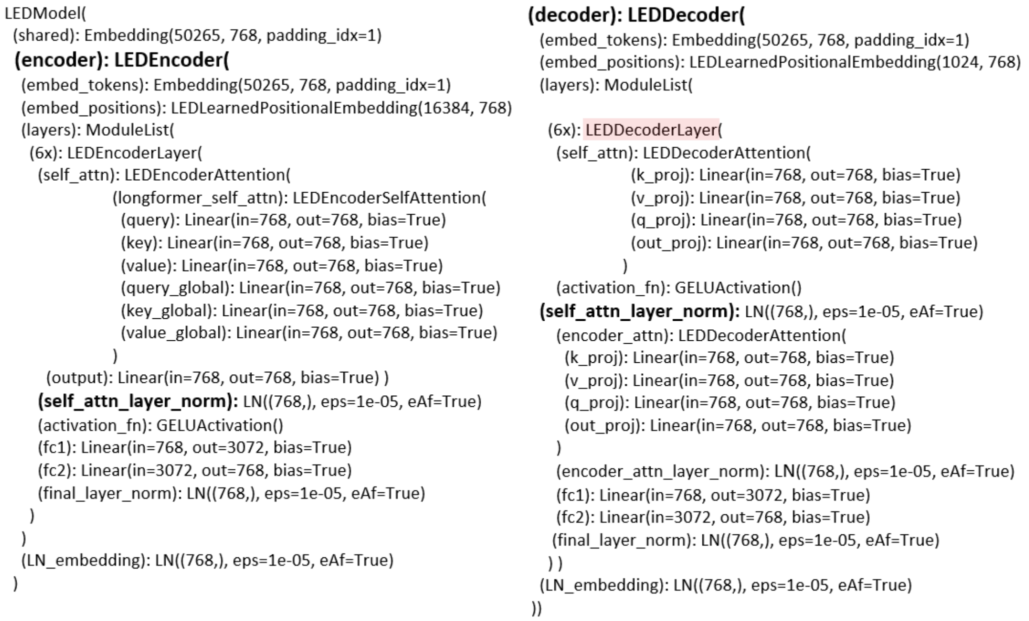


Figure A.4: Exploring Attention in Trained Longformer-3


```
LEDModel(
  (shared): Embedding(50265, 768, padding_idx=1)
  (encoder): LEDEncoder(
   (embed_tokens): Embedding(50265, 768, padding_idx=1)
   (embed_positions): LEDLearnedPositionalEmbedding(16384, 768)
   (layers): ModuleList(
     (6x): LEDEncoderLayer(
      (self_attn): LEDEncoderAttention(
                 (longformer_self_attn): LEDEncoderSelfAttention(
                  (query): Linear(in=768, out=768, bias=True)
                  (key): Linear(in=768, out=768, bias=True)
                  (value): Linear(in=768, out=768, bias=True)
                  (query_global): Linear(in=768, out=768, bias=True)
                  (key_global): Linear(in=768, out=768, bias=True)
                  (value_global): Linear(in=768, out=768, bias=True)
                 )
       (output): Linear(in=768, out=768, bias=True) )
      (self_attn_layer_norm): LN((768,), eps=1e-05, eAf=True)
      (activation_fn): GELUActivation()
      (fc1): Linear(in=768, out=3072, bias=True)
      (fc2): Linear(in=3072, out=768, bias=True)
      (final_layer_norm): LN((768,), eps=1e-05, eAf=True)
    )
   )
   (LN_embedding): LN((768,), eps=1e-05, eAf=True)
  )
```

```
(decoder): LEDDecoder(
  (embed_tokens): Embedding(50265, 768, padding_idx=1)
  (embed_positions): LEDLearnedPositionalEmbedding(1024, 768)
  (layers): ModuleList(

   (6x): LEDDecoderLayer(
    (self_attn): LEDDecoderAttention(
               (k_proj): Linear(in=768, out=768, bias=True)
               (v_proj): Linear(in=768, out=768, bias=True)
               (q_proj): Linear(in=768, out=768, bias=True)
               (out_proj): Linear(in=768, out=768, bias=True)
              )
    (activation_fn): GELUActivation()
  (self_attn_layer_norm): LN((768,), eps=1e-05, eAf=True)
    (encoder_attn): LEDDecoderAttention(
     (k_proj): Linear(in=768, out=768, bias=True)
     (v_proj): Linear(in=768, out=768, bias=True)
     (q_proj): Linear(in=768, out=768, bias=True)
     (out_proj): Linear(in=768, out=768, bias=True)
    )
    (encoder_attn_layer_norm): LN((768,), eps=1e-05, eAf=True)
    (fc1): Linear(in=768, out=3072, bias=True)
    (fc2): Linear(in=3072, out=768, bias=True)
   (final_layer_norm): LN((768,), eps=1e-05, eAf=True)
  ) )
  (LN_embedding): LN((768,), eps=1e-05, eAf=True)
))
```

Figure A.5: LED (Longformer Encoder Decoder) Model Architecture

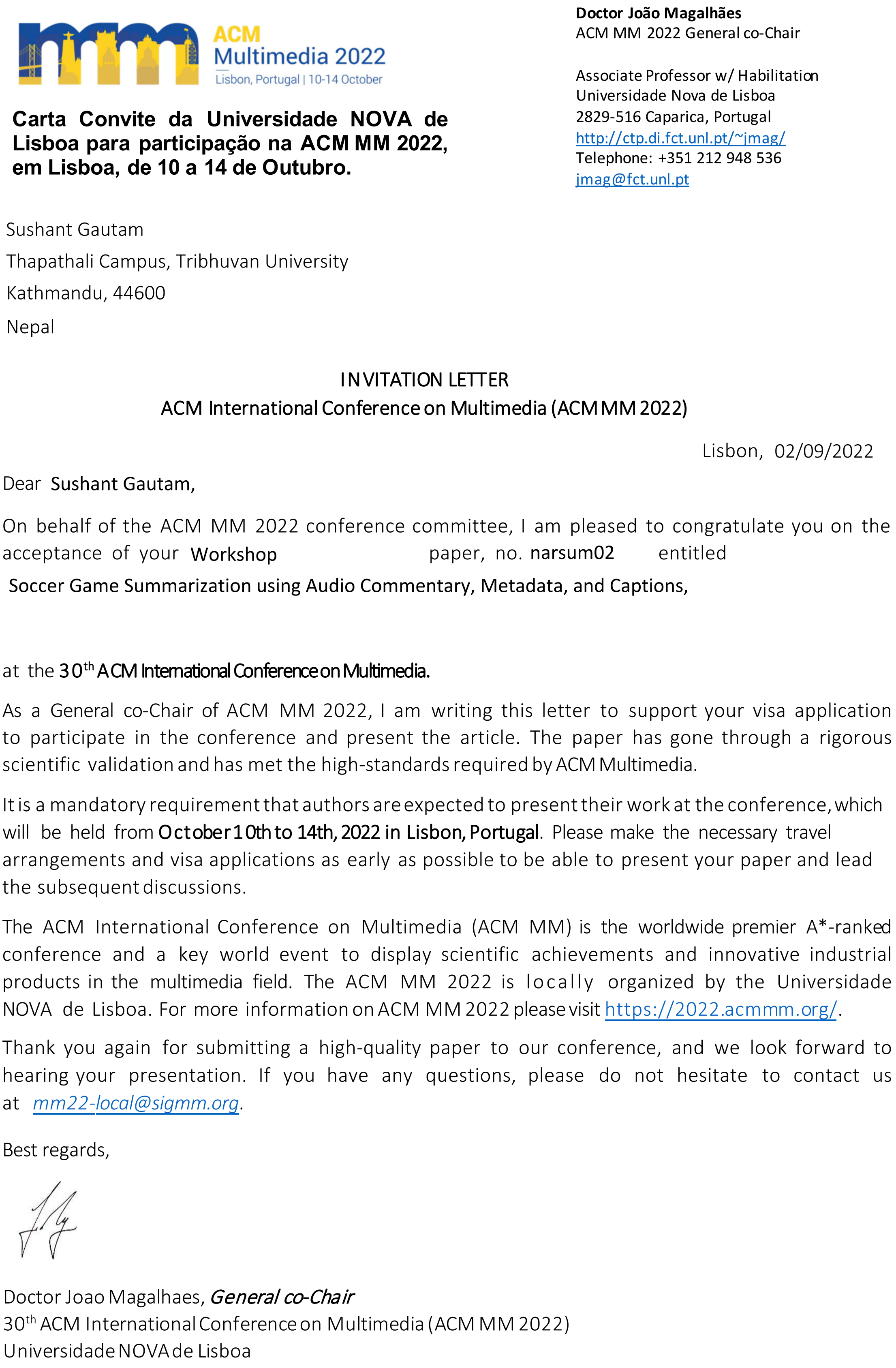
A.7 Associated Publication

1. S. Gautam, C. Midoglu, S. S. Sabet, D. B. Kshatri, and P. Halvorsen, “Soccer Game Summarization using Audio Commentary, Metadata, and Captions,” in *Proceedings of the 1st Workshop on User-centric Narrative Summarization of Long Videos (NarSUM ’22), October 10*, 2022, 2022, p. 10. doi: 10.1145/3552463.3557019. [106]

2. S. Gautam, C. Midoglu, S. S. Sabet, D. B. Kshatri, and P. Halvorsen. "Assisting Soccer Game Summarization via Audio Intensity Analysis of Game Highlights." *ResearchGate [Preprint.]*, 22 Sept. 2022. doi:10.13140/RG.2.2.34457.70240. [107]
   (Under-review, Submitted to *12th IOE Graduate Conference*)

The proof of acceptance or submission of the publications mentioned above is included in the upcoming pages.

A.7.1 Publication Acceptance Letter



ACM
Multimedia 2022
Lisbon, Portugal | 10-14 October

Carta Convite da Universidade NOVA de Lisboa para participação na ACM MM 2022, em Lisboa, de 10 a 14 de Outubro.

Doctor João Magalhães
ACM MM 2022 General co-Chair

Associate Professor w/ Habilitation
Universidade Nova de Lisboa
2829-516 Caparica, Portugal
http://ctp.di.fct.unl.pt/~jmag/
Telephone: +351 212 948 536
jmag@fct.unl.pt

Sushant Gautam
Thapathali Campus, Tribhuvan University
Kathmandu, 44600
Nepal

INVITATION LETTER
ACM International Conference on Multimedia (ACM MM 2022)

Lisbon, 02/09/2022

Dear Sushant Gautam,

On behalf of the ACM MM 2022 conference committee, I am pleased to congratulate you on the acceptance of your Workshop paper, no. narsum02 entitled

Soccer Game Summarization using Audio Commentary, Metadata, and Captions,

at the 30th **ACM International Conference on Multimedia.**

As a General co-Chair of ACM MM 2022, I am writing this letter to support your visa application to participate in the conference and present the article. The paper has gone through a rigorous scientific validation and has met the high-standards required by ACM Multimedia.

It is a mandatory requirement that authors are expected to present their work at the conference, which will be held from **October 10th to 14th, 2022 in Lisbon, Portugal**. Please make the necessary travel arrangements and visa applications as early as possible to be able to present your paper and lead the subsequent discussions.

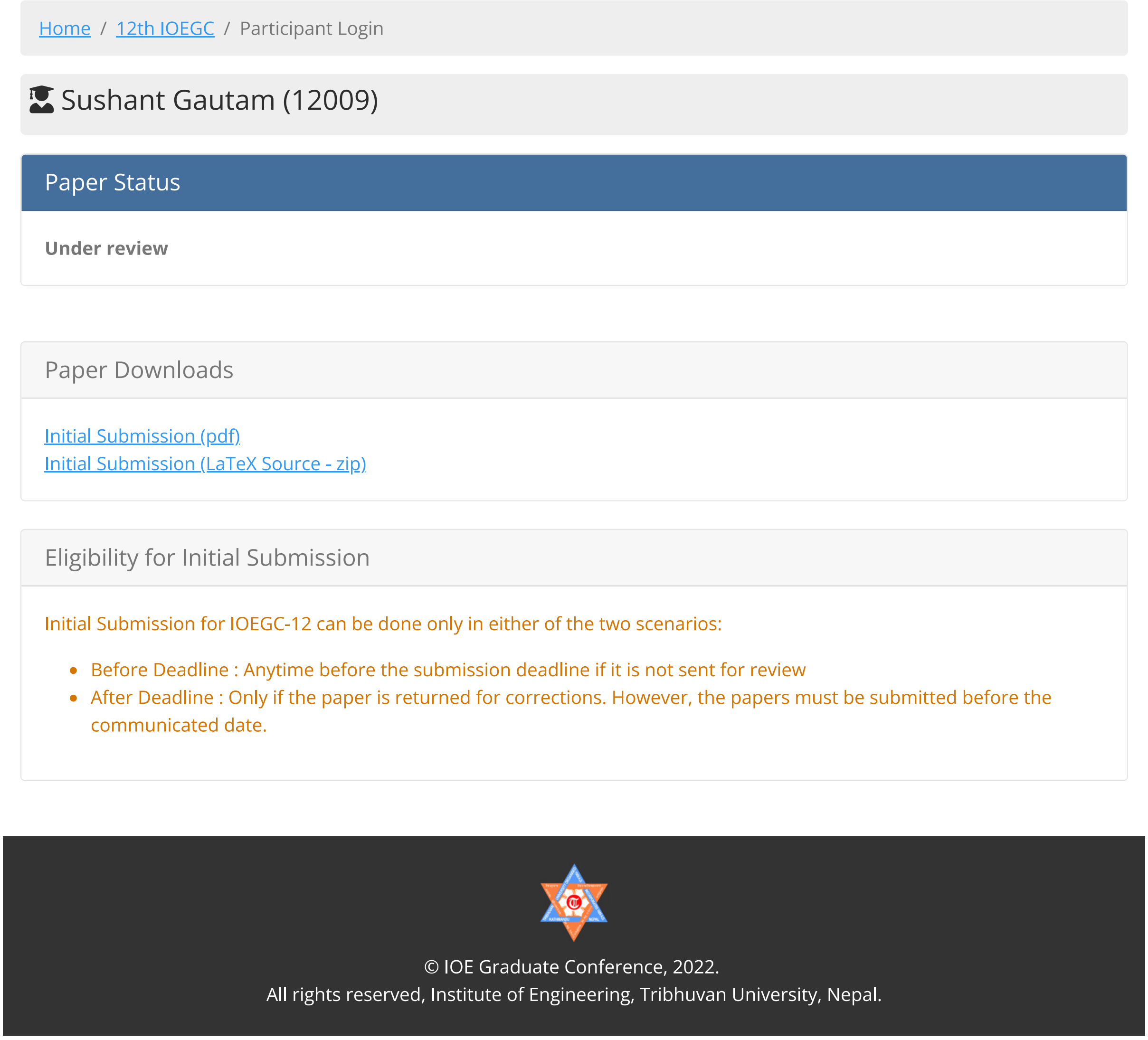
The ACM International Conference on Multimedia (ACM MM) is the worldwide premier A*-ranked conference and a key world event to display scientific achievements and innovative industrial products in the multimedia field. The ACM MM 2022 is locally organized by the Universidade NOVA de Lisboa. For more information on ACM MM 2022 please visit https://2022.acmmm.org/.

Thank you again for submitting a high-quality paper to our conference, and we look forward to hearing your presentation. If you have any questions, please do not hesitate to contact us at *mm22-local@sigmm.org*.

Best regards,

Doctor Joao Magalhaes, ***General co-Chair***
30th ACM International Conference on Multimedia (ACM MM 2022)
Universidade NOVA de Lisboa

A.7.2 Proof of Manuscript Submission



Home / 12th IOEGC / Participant Login

Sushant Gautam (12009)

Paper Status

Under review

Paper Downloads

Initial Submission (pdf)
Initial Submission (LaTeX Source - zip)

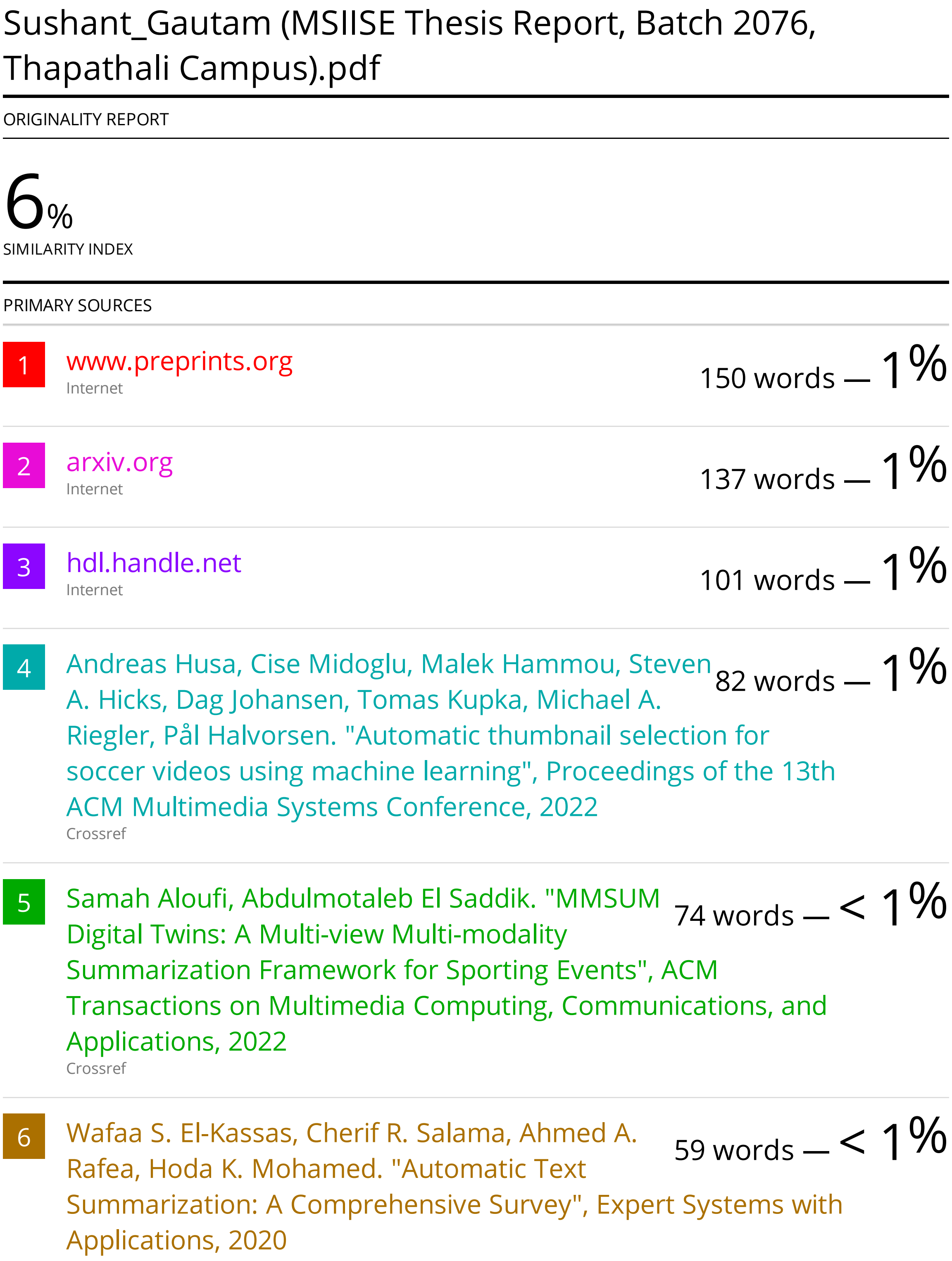
Eligibility for Initial Submission

Initial Submission for IOEGC-12 can be done only in either of the two scenarios:

- Before Deadline : Anytime before the submission deadline if it is not sent for review
- After Deadline : Only if the paper is returned for corrections. However, the papers must be submitted before the communicated date.

© IOE Graduate Conference, 2022.
All rights reserved, Institute of Engineering, Tribhuvan University, Nepal.

A.8 Plagiarism Report



Sushant_Gautam (MSIISE Thesis Report, Batch 2076, Thapathali Campus).pdf

ORIGINALITY REPORT

6%

SIMILARITY INDEX

PRIMARY SOURCES

1	www.preprints.org Internet	150 words — 1%
2	arxiv.org Internet	137 words — 1%
3	hdl.handle.net Internet	101 words — 1%
4	Andreas Husa, Cise Midoglu, Malek Hammou, Steven A. Hicks, Dag Johansen, Tomas Kupka, Michael A. Riegler, Pål Halvorsen. "Automatic thumbnail selection for soccer videos using machine learning", Proceedings of the 13th ACM Multimedia Systems Conference, 2022 Crossref	82 words — 1%
5	Samah Aloufi, Abdulmotaleb El Saddik. "MMSUM Digital Twins: A Multi-view Multi-modality Summarization Framework for Sporting Events", ACM Transactions on Multimedia Computing, Communications, and Applications, 2022 Crossref	74 words — < 1%
6	Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, Hoda K. Mohamed. "Automatic Text Summarization: A Comprehensive Survey", Expert Systems with Applications, 2020	59 words — < 1%

	Crossref	
7	www.scribbr.com Internet	56 words — < 1%
8	booksc.org Internet	34 words — < 1%
9	www.researchgate.net Internet	20 words — < 1%
10	Zheng, Chujie. "Abstractive Text Summarization via Contextual Semantics Understanding", University of Delaware, 2022 ProQuest	18 words — < 1%
11	Nada A. Dief, Ali E. Al-Desouky, Amr Aly Eldin, Asmaa M. El-Said. "An Adaptive Semantic Descriptive Model for Multi-Document Representation to Enhance Generic Summarization", International Journal of Software Engineering and Knowledge Engineering, 2017 Crossref	16 words — < 1%
12	Bo Feng. "Possible effect of solar activity on variation of the tree-rings of a 500 a platycladus orientalis at the Mausoleum of Emperor Huang", Science in China Series G Physics Mechanics and Astronomy, 04/2009 Crossref	13 words — < 1%
13	export.arxiv.org Internet	12 words — < 1%
14	Syed, Munira. "Learning and Inferring User Characteristics From Online Behavior and Content", University of Notre Dame, 2022 ProQuest	11 words — < 1%
15	acims.asu.edu	

	Internet	11 words — < 1%
16	mmsys2022.ie Internet	11 words — < 1%
17	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2019 Crossref	9 words — < 1%
18	Faiz R. Fajary, Tri W. Hadi, Shigeo Yoden. "Contributing Factors to Spatiotemporal Variations of Outgoing Longwave Radiation (OLR) in the Tropics", Journal of Climate, 2019 Crossref	9 words — < 1%
19	Schlosser, Jeffrey Steven. "Robotic Ultrasound Image Guidance for Radiation Therapy.", Stanford University, 2020 ProQuest	9 words — < 1%
20	dokumen.pub Internet	9 words — < 1%
21	"Machine Translation", Springer Science and Business Media LLC, 2016 Crossref	8 words — < 1%
22	Ercan, Gonenc. "Lexical Cohesion Analysis for Topic Segmentation, Summarization and Keyphrase Extraction", Bilkent Universitesi (Turkey) ProQuest	8 words — < 1%
23	Gonenc Ercan. "Lexical Cohesion Based Topic Modeling for Summarization", Lecture Notes in Computer Science, 2008 Crossref	8 words — < 1%

24	Snedden, Gregg. "River, tidal, and wind interactions in a deltaic estuarine system", Proquest, 20111109 ProQuest	8 words — < 1%
25	mafiadoc.com Internet	8 words — < 1%
26	www.educative.io Internet	8 words — < 1%
27	www.ieeesem.com Internet	8 words — < 1%
28	M. Kutlu. "Generic Text Summarization for Turkish", The Computer Journal, 01/06/2010 Crossref	6 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF

REFERENCES

[1] Vimeo Livestream Blog. Streaming stats - 47 must-know live video streaming statistics. `https://livestream.com/blog/62-must-know-stats-live-video-streaming`, 2022.

[2] FIFA.com. More than half the world watched record-breaking 2018 world cup, 2018.

[3] Torrens University Australia. Why the sports industry is booming in 2020 (and which key players are driving growth), 2020.

[4] Cise Midoglu, Steven A. Hicks, Vajira Thambawita, Tomas Kupka, and Pål Halvorsen. Mmsys'22 grand challenge on ai-based video production for soccer. In *13th ACM Multimedia Systems Conference (MMSys'22)*. ACM, 2022.

[5] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, 2014.

[6] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, page 568–576, 2014.

[7] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1933–1941, 2016.

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[9] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1049–1058, 2016.

[10] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference Computer Vision (ECCV)*, pages 20–36, 2016.

[11] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, page 4489–4497, 2015.

[12] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6459, 2018.

[13] H. Idrees, A. R. Zamir, Y. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. The thumos challenge on action recognition for videos “in the wild”. *Computer Vision and Image Understanding*, 155:1–23, 2017.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.

[15] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1711–1721, June 2018.

[16] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017.

[17] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[18] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference Computer Vision (ECCV)*, 2018.

[19] Olav A. Nergård Rongved, Steven A. Hicks, Vajira Thambawita, Håkon K. Stensland, Evi Zouganeli, Dag Johansen, Michael A. Riegler, and Pål Halvorsen. Real-time detection of events in soccer videos using 3d convolutional neural networks. In *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, pages 135–144, 2020.

[20] Harilaos Koumaras, Georgios Gardikis, George Xilouris, Evangelos Pallis, and Anastasios Kourtis. Shot boundary detection without threshold parameters. *J. Electronic Imaging*, 15:020503, 4 2006.

[21] Hossam M. Zawbaa, Nashwa El-Bendary, Aboul Ella Hassanien, and Ajith Abraham. Svm-based soccer video summarization system. In *Proceedings of the World Congress on Nature and Biologically Inspired Computing*, pages 7–11, 2011.

[22] Hossam Zawbaa, Nashwa El-Bendary, Aboul Ella Hassanien, and Tai-Hoon Kim. Event detection based approach for soccer video summarization using machine learning. *International Journal of Multimedia and Ubiquitous Engineering (IJ-MUE)*, 7, 1 2012.

[23] Peng Xu, Lexing Xie, Shih-Fu Chang, A. Divakaran, A. Vetro, and Huifang Sun. Algorithms and system for segmentation and structure analysis in soccer video. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, pages 721–724, 2001.

[24] Muhammad Rafiq, Ghazala Rafiq, Rockson Agyeman, Seong-Il Jin, and Gyu Sang Choi. Scene classification for sports video summarization using transfer learning. *Sensors*, 20:1702, 03 2020.

[25] Reede Ren and Joemon M. Jose. Football video segmentation based on video production strategy. In *Proceedings of ECIR - Advances in Information Retrieval*, pages 433–446, 2005.

[26] Dian Tjondronegoro, Yi-Ping Phoebe Chen, and Binh Pham. Sports video summarization using highlights and play-breaks. In *Proceedings of ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR)*, page 201–208, 2003.

[27] Arnau Raventos, Raul Quijada, Luis Torres, and Francesc Tarres. Automatic summarization of soccer highlights using audio-visual descriptors, 2014.

[28] Chen-Yu Chen, Jia-Ching Wang, Jhing-Fa Wang, and Yu-Hen Hu. Motion entropy feature and its applications to event-based segmentation of sports video. *EURASIP Journal on Advances in Signal Processing*, 2008, 2008.

[29] Johannes Kopf, Michael F Cohen, and Richard Szeliski. First-person hyper-lapse videos. *ACM Transactions on Graphics (TOG)*, 33(4):1–10, 2014.

[30] Bin Zhao and Eric P. Xing. Quasi Real-Time Summarization for Consumer Videos. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2513–2520. IEEE, June 2014.

[31] Yair Poleg, Chetan Arora, and Shmuel Peleg. Temporal Segmentation of Egocentric Videos. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2544. IEEE, June 2014.

[32] Z. Rasheed and M. Shah. Scene detection in Hollywood movies and TV shows. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–343. IEEE, June 2003.

[33] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-Specific Video Summarization. In *Computer Vision – ECCV 2014*, pages 540–555. Springer, Cham, Switzerland, 2014.

[34] Amy Pavel, Björn Hartmann, and Maneesh Agrawala. Shot Orientation Controls for Interactive Cinematography with 360 Video. In *UIST '17: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pages 289–297. Association for Computing Machinery, New York, NY, USA, October 2017.

[35] Abdelati Malek Amel, Ben Abdelali Abdessalem, and Mtibaa Abdellatif. Video shot boundary detection using motion activity descriptor. *arXiv*, April 2010.

[36] Jiebo Luo, Christophe Papin, and Kathleen Costello. Towards Extracting Semantically Meaningful Key Frames From Personal Video Clips: From Humans to Computers. *IEEE Trans. Circuits Syst. Video Technol.*, 19(2):289–301, December 2008.

[37] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video Summarization with Long Short-Term Memory. In *Computer Vision – ECCV 2016*, pages 766–782. Springer, Cham, Switzerland, September 2016.

[38] Feng Zhou, Sing Bing Kang, and Michael F. Cohen. Time-Mapping Using Space-Time Saliency. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3358–3365. IEEE, June 2014.

[39] M. Sun, Ali Farhadi, and S. Seitz. Ranking Domain-Specific Highlights by Analyzing Edited Videos. *undefined*, 2014.

[40] Samy Bengio, Fernando Pereira, Yoram Singer, and Dennis Strelow. Group Sparse Coding. *Advances in Neural Information Processing Systems*, 22, 2009.

[41] Jiatong Li, Ting Yao, Qiang Ling, and Tao Mei. Detecting shot boundary with sparse coding for video summarization. *Neurocomputing*, 266:66–78, 2017.

[42] Ping Li, Yanwen Guo, and Hanqiu Sun. Multi-keyframe abstraction from videos. *ResearchGate*, pages 2473–2476, September 2011.

[43] Rafik Hamza, Khan Muhammad, Zhihan Lv, and Faiza Titouna. Secure video summarization framework for personalized wireless capsule endoscopy. *Pervasive Mob. Comput.*, 41:436–450, October 2017.

[44] Dan B. Goldman, Brian Curless, David Salesin, and Steven M. Seitz. Schematic storyboarding for video visualization and editing. *ACM Trans. Graphics*, 25(3):862–871, July 2006.

[45] Ba Tu Truong and Svetha Venkatesh. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(1):3–es, February 2007.

[46] Amy Pavel, Dan B. Goldman, Björn Hartmann, and Maneesh Agrawala. SceneSkim: Searching and Browsing Movies Using Synchronized Captions, Scripts and Plot Summaries. In *UIST '15: Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 181–190. Association for Computing Machinery, New York, NY, USA, November 2015.

[47] Yang Cong, Junsong Yuan, and Jiebo Luo. Towards Scalable Summarization of Consumer Videos Via Sparse Dictionary Selection. *IEEE Trans. Multimedia*, 14(1):66–75, September 2011.

[48] Min Sun, Ali Farhadi, Ben Taskar, and Steve Seitz. Salient Montages from Unconstrained Videos. In *Computer Vision – ECCV 2014*, pages 472–488. Springer, Cham, Switzerland, 2014.

[49] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, pages 2069–2077. MIT Press, Cambridge, MA, USA, December 2014.

[50] Peter Mindek, Ladislav Čmolík, Ivan Viola, Eduard Gröller, and Stefan Bruckner. Automatized summarization of multiplayer games. In *SCCG '15: Proceedings of the 31st Spring Conference on Computer Graphics*, pages 73–80. Association for Computing Machinery, New York, NY, USA, April 2015.

[51] M. U. Sreeja and Binsu C. Kovoor. Towards genre-specific frameworks for video summarisation: A survey. *J. Visual Commun. Image Represent.*, 62:340–358, July 2019.

[52] Suppawong Tuarob, Sumit Bhatia, Prasenjit Mitra, and C. Lee Giles. AlgorithmSeer: A System for Extracting and Searching for Algorithms in Scholarly Big Data. *IEEE Trans. Big Data*, 2(1):3–17, April 2016.

[53] Evi Yulianti, Ruey-Cheng Chen, Falk Scholer, W. Bruce Croft, and Mark Sanderson. Document Summarization for Answering Non-Factoid Queries. *IEEE Trans. Knowl. Data Eng.*, 30(1):15–28, September 2017.

[54] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video. *IEEE Trans. Knowl. Data Eng.*, 31(5):996–1009, June 2018.

[55] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, December 2021.

[56] Chris van der Lee, Emiel Krahmer, and Sander Wubben. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. *ACL Anthology*, pages 95–104, September 2017.

[57] Nadine Braun, Martijn Goudbeek, and Emiel Krahmer. The Multilingual Affective Soccer Corpus (MASC): Compiling a biased parallel corpus on soccer reportage in English, German and Dutch. *ACL Anthology*, pages 74–78, 2016.

[58] Lorenzo Gatti, Chris van der Lee, and Mariët Theune. Template-based multilingual football reports generation using Wikidata as a knowledge base. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 183–188. Association for Computational Linguistics (ACL), 2018.

[59] David PA Corney, Carlos Martin, and Ayse Göker. Two sides to every story: Subjective event summarization of sports events using twitter. In *SoMuS@ ICMR*. Citeseer, 2014.

[60] David L. Chen and Raymond J. Mooney. Learning to sportscast: a test of grounded language acquisition. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 128–135. Association for Computing Machinery, New York, NY, USA, July 2008.

[61] Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. Data-Driven News Generation for Automated Journalism. *ACL Anthology*, pages 188–197, September 2017.

[62] Vassilis Plachouras, Charese Smiley, Hiroko Bretz, Ola Taylor, Jochen L. Leidner, Dezhao Song, and Frank Schilder. Interacting with Financial Data using Natural Language. In *SIGIR '16: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1121–1124. Association for Computing Machinery, New York, NY, USA, July 2016.

[63] Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. Choosing words in computer-generated weather forecasts. *Artif. Intell.*, 167(1):137–169, September 2005.

[64] Mary Lynn Young and Alfred Hermida. From Mr. and Mrs. Outlier To Central Tendencies. *Digital Journalism*, 3(3):381–397, May 2015.

[65] Will Oremus. The First News Report on the L.A. Earthquake Was Written by a Robot. *Slate Magazine*, March 2014.

[66] General election 2019: Semi-automation makes it a night of 689 stories, April 2019. [Online; accessed 17. May 2022].

[67] Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. Summarizing microblogs automatically. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, pages 685–688, 2010.

[68] David Inouye and Jugal K. Kalita. Comparing Twitter Summarization Algorithms for Multiple Post Summaries. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 298–306. IEEE, October 2011.

[69] Deepayan Chakrabarti and Kunal Punera. Event Summarization Using Tweets. *ICWSM*, 5(1):66–73, 2011.

[70] Freddy Chua and Sitaram Asur. Automatic Summarization of Events from Social Media. *ICWSM*, 7(1):81–90, 2013.

[71] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. Summarizing sporting events using twitter. In *IUI '12: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 189–198. Association for Computing Machinery, New York, NY, USA, February 2012.

[72] Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, and Julio Gonzalo. Towards real-time summarization of scheduled events from twitter streams. In *HT '12: Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 319–320. Association for Computing Machinery, New York, NY, USA, June 2012.

[73] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 227–236. Association for Computing Machinery, New York, NY, USA, May 2011.

[74] Mitsumasa Kubo, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Generating Live Sports Updates from Twitter by Finding Good Reporters. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 527–534. IEEE, November 2013.

[75] Liang-Chi Hsieh, Ching-Wei Lee, Tzu-Hsuan Chiu, and Winston Hsu. Live Semantic Sport Highlight Detection Based on Analyzing Tweets of Twitter. In *2012 IEEE International Conference on Multimedia and Expo*, pages 949–954. IEEE, July 2012.

[76] Sushant Gautam, Saloni Shikha, Alina Devkota, and Spandan Pyakurel. Sentence ranking and answer pinpointing in online discussion forums utilising user-generated metrics and highlights. In *Proceedings of the NaSCoIT 2018: 9th National Student's Conference on Information Technology and 4th International IT Conference on ICT with Smart Computing.*, Kathmandu, Nepal. ACM.

[77] Samah Aloufi and Abdulmotaleb El Saddik. MMSUM Digital Twins: A Multi-

view Multi-modality Summarization Framework for Sporting Events. *ACM Trans. Multimedia Comput. Commun. Appl.*, 18(1):1–25, January 2022.

[78] Baoxin Li, Hao Pan, and Ibrahim Sezan. A general framework for sports video summarization with its application to soccer. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 3, pages III–169. IEEE, 2003.

[79] Ali Javed, Khalid Bashir Bajwa, Hafiz Malik, Aun Irtaza, and Muhammad Tariq Mahmood. A hybrid approach for summarization of cricket videos. In *2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pages 1–4. IEEE, 2016.

[80] Antonio Tejero-de Pablos, Yuta Nakashima, Tomokazu Sato, Naokazu Yokoya, Marko Linna, and Esa Rahtu. Summarization of User-Generated Sports Video by Using Deep Action Recognition Features. *IEEE Trans. Multimedia*, 20(8):2000–2011, January 2018.

[81] Jiaan Wang, Zhixu Li, Tingyi Zhang, Duo Zheng, Jianfeng Qu, An Liu, Lei Zhao, and Zhigang Chen. Knowledge Enhanced Sports Game Summarization. In *WSDM '22: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1045–1053. Association for Computing Machinery, New York, NY, USA, February 2022.

[82] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[83] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR, 2017.

[84] John Cameron. Signals, Sound and Sensation, by William M. Hartmann. *Med. Phys.*, 25(2):256, February 1998.

[85] Christopher E Heil and David F Walnut. Continuous and discrete wavelet transforms. *SIAM review*, 31(4):628–666, 1989.

[86] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1792–179210. IEEE, June 2018.

[87] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4508–4519, June 2021.

[88] Kuan-Hao Huang, Chen Li, and Kai-Wei Chang. Generating sports news from live commentary: A chinese dataset for sports game summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL)*, 2020.

[89] Suramya Tomar. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10, 2006.

[90] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154, 2020. Deezer Research.

[91] Amazon Transcribe – Speech to Text - AWS, July 2022. [Online; accessed 23. Jul. 2022].

[92] Speech to Text – Audio to Text Translation | Microsoft Azure, July 2022. [Online; accessed 23. Jul. 2022].

[93] Speech-to-Text: Automatic Speech Recognition | Google Cloud, July 2022. [Online; accessed 23. Jul. 2022].

[94] IBM Watson - Speech to Text, August 2021. [Online; accessed 23. Jul. 2022].

[95] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[96] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer. *arXiv*, April 2020.

[97] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. *ACL Anthology*, pages 311–318, July 2002.

[98] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. *ACL Anthology*, pages 340–348, August 2010.

[99] Virendra Gupta and T. J. Siddiqui. *Multi-document summarization using sentence clustering*. December 2012.

[100] Elena Lloret, Laura Plaza, and Ahmet Aker. The challenging task of summary evaluation: an overview. *Lang. Resources &. Evaluation*, 52(1):101–148, March 2018.

[101] Shuai Wang, Xiang Zhao, Bo Li, Bin Ge, and Daquan Tang. Integrating Extractive and Abstractive Models for Long Text Summarization. In *2017 IEEE International Congress on Big Data (BigData Congress)*, pages 305–312. IEEE, June 2017.

[102] Rui Sun, Zhenchao Wang, Yafeng Ren, and Donghong Ji. Query-biased multi-document abstractive summarization via submodular maximization using event guidance. In *International Conference on Web-Age Information Management*, pages 310–322. Springer, 2016.

[103] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. A Survey of Evaluation Metrics Used for NLG Systems. *ACM Comput. Surv.*, 55(2):1–39, January 2022.

[104] Translator – Translation Software as a Service | Microsoft Azure, July 2022. [Online; accessed 23. Jul. 2022].

[105] Ruochen Zhang and Carsten Eickhoff. SOCCER: An Information-Sparse Discourse State Tracking Collection in the Sports Commentary Domain. *ACL Anthology*, pages 4325–4333, June 2021.

[106] Sushant Gautam, Cise Midoglu, Saeed Shafiee Sabet, Dinesh Baniya Kshatri, and Pål Halvorsen. Soccer game summarization using audio commentary, metadata, and captions. In *Proceedings of the 1st Workshop on User-centric Narrative Summarization of Long Videos (NarSUM '22), October 10, 2022*, page 10, Lisboa, Portugal. ACM.

[107] Sushant Gautam, Cise Midoglu, Saeed Shafiee Sabet, Dinesh Baniya Kshatri, and Pål Halvorsen. Assisting Soccer Game Summarization via Audio Intensity Analysis of Game Highlights. *ResearchGate [Preprint.]*, September 2022.