UiA University of Agder

Master

Deep learning-based golf swing sequencing from videos

Neha

SUPERVISOR
Morten Goodwin, Per-Arne Andersen, and Daniel Groos

University of Agder, 2024
Faculty of Engineering and Science
Department of Engineering and Sciences

Obligatorisk gruppeerklæring

Den enkelte student er selv ansvarlig for å sette seg inn i hva som er lovlige hjelpemidler, retningslinjer for bruk av disse og regler om kildebruk. Erklæringen skal bevisstgjøre studentene på deres ansvar og hvilke konsekvenser fusk kan medføre. Manglende erklæring fritar ikke studentene fra sitt ansvar.

1.	Vi erklærer herved at vår besvarelse er vårt eget arbeid, og at vi ikke har brukt andre kilder eller har mottatt annen hjelp enn det som er nevnt i besvarelsen.	Ja
2.	**Vi erklærer videre at denne besvarelsen:** • Ikke har vært brukt til annen eksamen ved annen avdeling/universitet/høgskole innenlands eller utenlands. • Ikke refererer til andres arbeid uten at det er oppgitt. • Ikke refererer til eget tidligere arbeid uten at det er oppgitt. • Har alle referansene oppgitt i litteraturlisten. • Ikke er en kopi, duplikat eller avskrift av andres arbeid eller besvarelse.	Ja
3.	Vi er kjent med at brudd på ovennevnte er å betrakte som fusk og kan medføre annullering av eksamen og utestengelse fra universiteter og høgskoler i Norge, jf. Universitets- og høgskoleloven §§4-7 og 4-8 og Forskrift om eksamen §§ 31.	Ja
4.	Vi er kjent med at alle innleverte oppgaver kan bli plagiatkontrollert.	Ja
5.	Vi er kjent med at Universitetet i Agder vil behandle alle saker hvor det forligger mistanke om fusk etter høgskolens retningslinjer for behandling av saker om fusk.	Ja
6.	Vi har satt oss inn i regler og retningslinjer i bruk av kilder og referanser på biblioteket sine nettsider.	Ja
7.	Vi har i flertall blitt enige om at innsatsen innad i gruppen er merkbart forskjellig og ønsker dermed å vurderes individuelt. Ordinært vurderes alle deltakere i prosjektet samlet.	Nei

Publiseringsavtale

Fullmakt til elektronisk publisering av oppgaven Forfatter(ne) har opphavsrett til oppgaven. Det betyr blant annet enerett til å gjøre verket tilgjengelig for allmennheten (Åndsverkloven. §2).
Oppgaver som er unntatt offentlighet eller taushetsbelagt/konfidensiell vil ikke bli publisert.

Vi gir herved Universitetet i Agder en vederlagsfri rett til å gjøre oppgaven tilgjengelig for elektronisk publisering:	Ja
Er oppgaven båndlagt (konfidensiell)?	Nei
Er oppgaven unntatt offentlighet?	Nei

Acknowledgements

I express my deepest gratitude to my academic supervisors, Prof. Morten Goodwin and Prof. Per-Arne Andersen from the University of Agder. Their expert guidance, patience, and invaluable advice have been essential to completing this thesis. Their willingness to share their vast knowledge and insightful feedback has profoundly shaped this research.

Special thanks to Daniel Groos from Initial Force AS, who provided expert advice and practical insights crucial to this project's success. His involvement is a testament to the impactful collaboration between academia and industry.

I am profoundly grateful to my family for their unending support throughout my studies. To my husband, Micheal Dutt, thank you for your unwavering support and encouragement. Your belief in my abilities has been a source of motivation and strength. To my son, Krithvik Dutt, who was born just before the start of my master's journey, you have been my greatest joy and inspiration. Watching you grow as I navigated this phase of my academic career has been the brightest part of the experience.

I am grateful for the assistance of digital tools that have facilitated my writing and research processes. Finally, I would like to thank all my peers and colleagues who have provided feedback and support, helping me refine my ideas and research.

Abstract

This thesis delves into the effectiveness of advanced deep learning configurations for identifying the dynamic and static phases of golf swings, a fundamental skill in golf that directly influences performance outcomes. Traditional deep learning models often struggle with detecting static movements, which are subtle but crucial for comprehensive motion analysis in sports. This research gap underscores a significant need for enhanced model architectures incorporating advanced deep learning techniques designed specifically for the complexity of sports motion analytics.

To address this challenge, the study explores four innovative deep learning configurations: MobileNetV2 + LSTM, ResNet50 + LSTM, MobileNetV3 + LSTM, and a novel integration of MobileNetV3 with a Convolutional Block Attention Module (CBAM + LSTM). Each configuration is rigorously tested to evaluate its proficiency in capturing the golf swing's pronounced and subtle movement. The experiments are structured to systematically assess and compare each model's ability to accurately detect phases of the swing, focusing on integrating spatial and temporal data critical for dynamic and static phase recognition.

The results demonstrate that while traditional configurations like MobileNetV2 + LSTM provide a solid foundation for detecting dynamic movements, they fail to capture static phases accurately. However, integrating CBAM with MobileNetV3 significantly enhances model performance, particularly detecting static phases. This improvement highlights the transformative potential of attention mechanisms in refining the focus and sensitivity of neural networks, enabling them to excel where conventional architectures falter. This research has profound implications. It offers a deeper understanding of the application of neural network architectures in sports analytics and paves the way for future advancements in automated coaching tools. By enhancing phase detection accuracy, this work contributes to developing more sophisticated analytics tools to provide athletes and coaches with precise, real-time feedback essential for performance optimization.

Keywords: Attention Mechanisms, Convolutional Block Attention Module (CBAM), Deep Learning, Golf Swing Analysis, Long Short-Term Memory (LSTM), Sports Analytics, Transfer Learning.

Contents

Acknowledgements ii

Abstract iii

List of Figures vii

List of Tables ix

1 Introduction 1
- 1.1 Motivation and Research Gap . . . 2
  - 1.1.1 Addressing the Research Gap . . . 2
  - 1.1.2 Implications for Sports Performance Technology . . . 3
- 1.2 Research Questions and Hypotheses . . . 3
- 1.3 Thesis Structure . . . 3

2 Literature Review 5
- 2.1 Human Activity Recognition Based on Video Datasets . . . 5
- 2.2 Deep Learning in Sports Analytics . . . 6
- 2.3 Deep Learning in Golf Sports . . . 8

3 Background 10
- 3.1 Convolution Neural Network (CNN) . . . 11
  - 3.1.1 Input Layer . . . 11
  - 3.1.2 Convolution Layer . . . 11
  - 3.1.3 ReLU Layer . . . 12
  - 3.1.4 Pooling Layer . . . 12
  - 3.1.5 Fully Connected Layer . . . 12
  - 3.1.6 Output Layer . . . 12
- 3.2 Convolutional Block Attention Module (CBAM) . . . 12
  - 3.2.1 Overview . . . 13
  - 3.2.2 Channel Attention Module (CAM) . . . 13
  - 3.2.3 Spatial Attention Module (SAM) . . . 14
  - 3.2.4 Integration into CNNs . . . 14
- 3.3 Recurrent Neural Networks and Long Short-Term Memory . . . 14
  - 3.3.1 Recurrent Neural Networks (RNNs) . . . 14
  - 3.3.2 Long Short-Term Memory (LSTM) Networks . . . 15
  - 3.3.3 Integration into CNNs with CBAM . . . 16
- 3.4 Advanced Neural Network Architectures . . . 16
  - 3.4.1 ResNet50 . . . 16
  - 3.4.2 MobileNetV2 . . . 18
  - 3.4.3 MobileNetV3 . . . 19

4 GolfDB Dataset **21**
4.1 Golf Swing Events . . . 21
4.2 Golf Swing Sequencing . . . 22
4.3 Annotation . . . 23

5 Methodology **24**
5.1 Data Prepocessing . . . 24
5.2 Data Loader . . . 25
5.3 Proposed Architecture . . . 26
5.3.1 Input . . . 28
5.3.2 Feature Extraction Module . . . 28
5.3.3 Feature Refinement Module . . . 28
5.3.4 Temporal Module . . . 29
5.3.5 Output . . . 29
5.4 Model Summary . . . 30
5.5 Model Training . . . 30
5.5.1 Training Configuration . . . 30
5.5.2 Computational Resource Allocation . . . 31
5.5.3 Batch Formulation . . . 31
5.5.4 Model Architecture Initialization . . . 31
5.5.5 Data Standardization . . . 32
5.5.6 Optimization Framework . . . 32
5.5.7 State Preservation and Learning Progression . . . 32
5.5.8 Training Iterations . . . 32
5.6 Model Evaluation . . . 33
5.6.1 Evaluation Configuration . . . 33
5.6.2 Computational Resource Allocation . . . 33
5.6.3 Evaluation Metrics . . . 33
5.6.4 Performance Assessment . . . 33
5.6.5 Model Generalization . . . 34

6 Results and Discussions **35**
6.1 MobileNetV2 + LSTM . . . 35
6.1.1 Experimental Setup . . . 35
6.1.2 Implementation Details . . . 35
6.1.3 Results . . . 36
6.1.4 Discussion . . . 36
6.2 ResNet50 + LSTM . . . 37
6.2.1 Experimental Setup . . . 37
6.2.2 Implementation Details . . . 37
6.2.3 Results . . . 37
6.2.4 Discussion . . . 38
6.3 MobileNetV3 + LSTM . . . 39
6.3.1 Experimental Setup . . . 39
6.3.2 Implementation Details . . . 39
6.3.3 Results . . . 39
6.3.4 Discussion . . . 39
6.4 MobileNetV3 + CBAM + LSTM . . . 40
6.4.1 Experimental Setup . . . 41
6.4.2 Implementation Details . . . 41
6.4.3 Results . . . 41
6.4.4 Discussion . . . 42
6.5 Discussion . . . 42

6.5.1 Research Question and Hypotheses Revisited 44
6.6 Comparative Performance Analysis of Our Approach Against SwingNet . . . 45

7 Conclusion and Future Work **47**
7.1 Conclusion 47
7.2 Future Work 48

Bibliography **49**

List of Figures

3.1 Basic convolution neural networks architecture. 11
3.2 Overview of convolutional block attention module. 13
3.3 Overview of Bi-directional Recurrent Neural Network. 15
3.4 Block diagram of Resnet-50 17
3.5 Overview of MobileNetV2 Convolutional Blocks 18
3.6 MobileNetV3 (MobileNetV2 + Squeeze-and-Excite) 20

4.1 Swing Events from one of the subjects from the GolfDB dataset. 22

5.1 Schematic representation of the proposed event detector neural network architecture. The architecture integrates a MobileNetV3-based Feature Extraction Module with Channel and Spatial Attention mechanisms for feature refinement, followed by a Bidirectional LSTM Temporal Module for capturing dynamic temporal dependencies. 27

6.1 Accuracy per Phase for Swing Net (MobileNetv2 + LSTM) 36
6.2 Accuracy per Phase for ResNet50 + LSTM 38
6.3 Accuracy per Phase for MobileNetV3 + LSTM 40
6.4 Accuracy per Phase for MobileNetV3 + CBAM + LSTM 41
6.5 Comparative accuracy of different models across various phases of the golf swing. 44

List of Tables

5.1 Detailed Summary of the EventDetector Model Architecture 30

6.1 Comparison of Different Models 46

Chapter 1

Introduction

Often considered a sport of finesse and precision, golf embodies a harmonious blend of strategic thinking, physical prowess, and a profound understanding of mechanics [47]. A single stroke's quality can influence the trajectory of a game, making precision and control critical elements of successful play. The golf swing, a fundamental skill in this sport, is a complex movement that determines the effectiveness of each stroke [31]. This motion involves multiple phases, including setup, backswing, downswing, impact, and follow-through; each plays a critical role in the performance outcome [25].

Golf's origins can be traced back to Scotland in the 15th century [16]. Since then, it has evolved into a global sport with millions of enthusiasts and a solid professional and amateur presence. Over the centuries, the understanding of golf mechanics has deepened, with players constantly seeking ways to refine their technique to gain competitive advantages [9]. This ongoing pursuit has fostered an environment ripe for technological integration, particularly in swing analysis, which directly influences performance improvement.

Traditionally, golf swing analysis has been the domain of experienced coaches who use their trained eyes to evaluate and improve players' techniques [51]. This form of analysis, while invaluable, carries inherent limitations—chiefly, its reliance on subjective assessment, which can lead to inconsistencies in coaching and player development. The introduction of high-speed cameras and motion capture technology marked significant progress [42], offering more detailed information on swing mechanics by capturing hundreds of frames per second. However, these technologies come with high costs and logistical demands, limiting their access to elite training facilities.

Despite their benefits, manual analysis and high-tech methods have drawbacks. One significant challenge is the inability to capture a golf swing's small yet critical details, particularly at high speeds [41]. Golf swings are rapid and complex movements, often completed in less than two seconds, making them challenging even the most sophisticated cameras during critical phases such as impact. This limitation can lead to gaps in data, potentially causing oversights in analysis and subsequent coaching strategies.

In addition, traditional methods require substantial equipment and investment from specialist personnel to operate and interpret the data, restricting these advanced analytics tools to well-funded sports institutes and professional players. This disparity raises the need for more accessible, accurate, and objective swing analysis methods that can benefit a broader range of players.

The advent of machine learning and significant advances in computer vision have opened new avenues for analyzing complex activities like golf swings [27]. Machine learning models, trained on sequential image data from videos, can automatically detect and label different

phases of a swing of golf [34]. This automation improves the accuracy and objectivity of the analysis and democratizes access to advanced swing analysis tools. Players and coaches can receive instant feedback on swing mechanics, enabling immediate adjustments and improvements without elaborate setups.

1.1 Motivation and Research Gap

The primary motivation for this research originates from a pressing need to enhance the accuracy, efficiency, and accessibility of golf swing analysis through advanced machine learning technologies. Traditional video analysis methods, which revolutionized performance coaching upon their introduction, are increasingly facing limitations that restrict their ability to accurately interpret the rapid, complex movements characteristic of a golf swing. These traditional methods often result in significant oversight during critical motion phases, potentially leading to misguided training and development strategies.

This thesis is conducted in collaboration with Initial Force AS, a pioneer in sports performance technology. The company's mission is to enhance athletic performance through accessible, innovative software and sensor solutions. Initial Force AS provides a wide array of video analysis tools and sensor integrations, making advanced data and analytics available to a broader audience beyond traditional university and sports laboratory settings.

Initial Force AS's offerings, which include platforms like Swing Catalyst and the developing Motion Catalyst, are designed to provide all-in-one solutions for sports analysis and improvement. The integration of the research outcomes from this thesis into Initial Force AS's technology stack aims to significantly enhance their software applications, equipping users with unparalleled real-time analysis and feedback capabilities. This collaboration not only ensures the practical application of the research findings but also positions the developed technologies to make a substantial impact in the field of sports analytics.

1.1.1 Addressing the Research Gap

Despite the advancements in video capture technology, the rapid motion of a golf swing, which is often completed in less than two seconds, poses unique challenges:

- High temporal resolution is required to capture these quick movements, a capability that conventional video analysis tools often lack.

- Subtle movements, which are critical for the accurate execution of various swing phases, are frequently lost amidst noise or are too subtle for simplistic computational methods to detect.

Conventional methods like manual observation or high-speed cameras typically fall short of providing the necessary real-time, detailed feedback that is crucial for effective performance coaching. These methods are also limited by the need for significant hardware investments, which can be a barrier to accessibility and widespread adoption.

This thesis aims to bridge these gaps by leveraging cutting-edge advancements in neural network architectures and attention mechanisms. These technologies are specifically tailored to enhance the detection of subtle and rapid movements within video data, offering a more sophisticated approach to analyzing complex motion dynamics.

1. **Improved Feature Extraction:** The research introduces methods designed to more accurately capture and analyze the swift, subtle, and complex features of a golf swing that are often overlooked by traditional techniques.

2. **Enhanced Temporal Analysis:** By providing a more detailed temporal breakdown and evaluation of golf swing phases, these methods facilitate a deeper understanding of the dynamics involved, enabling more effective coaching and performance improvement.

The research evaluates four advanced deep learning configurations: MobileNetV2 + LSTM, ResNet50 + LSTM, MobileNetV3 + LSTM, and MobileNetV3 integrated with a Convolutional Block Attention Module (CBAM + LSTM). Each configuration is assessed for its effectiveness in detecting both subtle and pronounced movements characteristic of golf swings, with a special emphasis on the improvements enabled by the integration of CBAM. This focus is particularly pertinent for addressing the challenges in detecting static phases of the swing, which are traditionally difficult to analyze due to their minimal movement and subtle dynamics.

1.1.2 Implications for Sports Performance Technology

The potential real-world impact of this research in the context of Initial Force AS's sports performance technology offerings is substantial. By enhancing golf swing analysis capabilities, the developed technologies can profoundly influence training regimes, providing athletes and coaches with more precise, actionable feedback. This advancement aligns with the broader objectives of sports analytics and human performance studies, promising benefits that extend beyond individual athletes to influence the field as a whole.

1.2 Research Questions and Hypotheses

Based on the identified research gaps, this thesis addresses a central research question, accompanied by three distinct hypotheses aimed at exploring the depths and impacts of advanced neural network architectures on golf swing analysis:
RQ1: How do advanced neural network architectures, especially those incorporating attention mechanisms, enhance golf swing analysis?

1. **H1:** Advanced neural network architectures improve the feature extraction capabilities from video data, leading to more accurate detection of golf swing phases.

2. **H2:** The incorporation of attention mechanisms significantly increases the precision of phase detection by focusing analysis on the most relevant features and minimizing the influence of background noise.

3. **H3:** Integrating these advanced technologies enhances the reliability and consistency of golf swing analysis across different environments and swing types.

1.3 Thesis Structure

This thesis is structured to build incrementally upon insights from each experiment, culminating in sophisticated models that integrate advanced neural network architectures and attention mechanisms. Here is an outline of each chapter:

- **Chapter 1: Introduction**
  This chapter sets the stage for the thesis, outlining the research's objectives, scope, and structure. It introduces the challenges and potential of using deep learning for sports analytics, specifically for analyzing golf swings.

- **Chapter 2: Literature Review**
  This chapter provides a comprehensive literature review, focusing on general human activity recognition, deep learning in sports analytics, and specific applications in golf sports.

- **Chapter 3: GolfDB Dataset**
  This chapter details the GolfDB dataset used for the experiments, including descriptions of golf swing events, swing sequencing, and the annotation process.

- **Chapter 4: Background**
  This chapter discusses the technical background necessary for understanding the models used in the thesis, including detailed discussions of CNNs, CBAM, and LSTMs.

- **Chapter 5: Proposed Methodology**
  This chapter outlines the proposed methodology for processing the GolfDB dataset, the architectural details of the models used, and the model training and evaluation strategies.

- **Chapter 6: Results and Discussions**
  This chapter presents the results of the tested model configurations and discusses the findings compared to previous models.

- **Chapter 7: Conclusion and Future Work**
  This chapter concludes the thesis by summarizing the findings and discussing their implications for sports analytics and coaching. Outline potential future research directions.

Chapter 2

Literature Review

This chapter provides a comprehensive review of the relevant literature in areas critical to the understanding and advancement of human activity recognition and its specific application in sports analytics. The focus is developing and evolving computational methods significantly influencing sports performance analysis, mainly through deep learning techniques. Each section discusses key contributions and highlights the current trends, challenges, and gaps that this research aims to address.

2.1 Human Activity Recognition Based on Video Datasets

Human activity recognition (HAR) from video datasets involves using advanced deep learning models to detect and classify human movements accurately. HAR's significance lies in its ability to interpret complex physical activities from digital video, making substantial contributions to areas ranging from security to sports analytics.

CNNs have dramatically improved the accuracy of HAR systems by effectively extracting robust features from video data. Ronao and Cho [45] demonstrated this by applying CNNs to distinguish between six distinct locomotion activities, notably outperforming traditional methods like MLP, Naive Bayes, and SVM. Further explorations into sensor placements on the body have shown how CNNs can adapt to variations in data acquisition, improving the model's performance by optimizing sensor locations [24].

Integrating traditional feature engineering with CNNs has also shown promising advancements. The HAR-Net model, which combines time and frequency domain features with CNN-extracted features, exemplifies how hybrid models can leverage the strengths of both approaches to improve activity classification [11]. Moreover, shallow CNN architectures have proven effective and computationally efficient for real-time applications on mobile devices, making HAR more accessible and feasible in everyday scenarios [44].

RNNs, especially LSTM networks, excel in handling the temporal sequences typical in HAR. They can capture long-term dependencies within time-series data, crucial for continuous and dynamic activity recognition [37]. Combining CNNs with RNNs in models like CNN-RNN has significantly enhanced HAR by synergistically processing spatial and temporal data [35].

Recent advancements have introduced models that focus on the most relevant features for activity classification through attention mechanisms, significantly boosting the accuracy and efficiency of HAR systems [54]. Additionally, employing transfer learning has enabled HAR systems to utilize pre-trained models, thereby improving performance across various scenarios and datasets [18].

Developing specialized architectures and learning strategies, such as deep convolutional networks with partial and complete weight sharing, has shown significant improvements in recognizing complex activities [32]. Innovative approaches like cross-channel communication in CNNs have also emerged, enhancing the model's ability to isolate specific features from multi-sensor data, thus improving recognition capabilities [23].

Advances in multimodal learning have allowed HAR systems to effectively utilize data from various sensors simultaneously, improving the system's robustness and adaptability [57]. The effective combination of CNNs and LSTMs highlights a comprehensive approach that harnesses spatial and temporal data for improved activity recognition [39].

The application of RNNs in HAR extends beyond traditional models, with newer variations like Independently RNN (IndRNN) [58] and Continuous Time RNN (CTRNN) [2] providing alternatives that tackle specific challenges in activity recognition more effectively.

The methodologies developed for general HAR are increasingly being adapted for sports analytics. The precise activity recognition enabled by these advanced models offers significant benefits for sports performance analysis, where accurate and real-time data is crucial for coaching and training. Techniques such as CNNs for spatial feature extraction and RNNs for sequence modeling are particularly beneficial in analyzing complex athletic movements and enhancing strategies based on predictive analytics. Integrating multimodal data into sports analytics, including video, sensor, and biomechanical data, further exemplifies how HAR technologies are revolutionizing the field, providing deeper insights and more effective training methods.

The continuous evolution of CNN and RNN applications in HAR enhances our capability to handle increasingly complex human activities. It paves the way for transformative advancements in sports analytics, pushing the boundaries of precision, efficiency, and effectiveness in athletic performance monitoring.

2.2 Deep Learning in Sports Analytics

Deep learning technologies have revolutionized sports analytics, providing sophisticated tools that improve the precision and depth of performance analysis. This section explores sensor-based and computer vision-based applications that significantly advance understanding and enhance sports performance.

Sensor-based technologies have transformed sports analytics by leveraging the power of IoT wearables to capture detailed performance metrics. These technologies enable continuous monitoring of athletes, providing real-time data crucial for assessing performance and health. Cooper et al. [7] underscored the importance of data reliability in sports performance systems by developing a statistical procedure to determine the reliability of such data, which is essential to distinguish between correct insights and errors. This approach has been pivotal in ensuring the accuracy of performance assessments.

Expanding on the utility of sensor-based applications, Hossain et al. [21] introduced the SoccerMate framework, a sophisticated model that utilizes restricted Boltzmann machines for evaluating soccer players' performances. This framework emphasizes the significance of individual performance metrics in sports analytics. Similarly, Ghosh, Ramamurthy, and Roy [14] looked at how different people perform in sports using a method based on K-nearest neighbors to measure stance errors between players, showing how difficult it is to generalize data between other people's performances.

To deal with the difficulties of combining different data types, Blank et al. [4] created real-time systems to sort IMU signals, making it easier to track and analyze physical activities as they happened. Additional notable contributions in this area include works by Steels et al. [49], who employed sensors to monitor detailed movements in badminton, providing crucial data for coaching and performance enhancement. Anik et al. [1] focused on classifying badminton strokes using sensor data to highlight how specific movements can be algorithmically identified and categorized.

Computer vision technologies offer equally transformative insights into sports analytics by enabling detailed visual data analysis. Thermal imaging techniques, discussed by Costello et al. [8] and Kirimtat et al. [28], provide valuable insights into the physiological demands of sports activities. They help track limb movements and identify stressed positions, crucial for enhancing player performance and preventing injuries.

Augmented and virtual reality technologies have been explored for their potential to improve training and performance. Bideau et al. [3] and Wu et al. [53] investigated virtual reality systems that simulate real-game environments, enhancing players' skills and understanding of game tactics through interactive experiences. These systems offer a controlled setting to experiment with different strategies and techniques, directly translating into improved athletic performance on the field.

Public datasets have spurred further advancements in sports analytics by enabling researchers to apply deep learning techniques to sports footage, enhancing object detection, tracking, and classification capabilities. FarajiDavar et al. cite farajidavar2011transductive used transductive transfer learning to improve action recognition accuracy. They changed the HOG3D features to make activity recognition in sports more robust. Furthermore, Vanderplaetse and Dupont [50] combined audio and visual data to improve the detection of soccer actions, demonstrating the effectiveness of multimodal systems in improving the precision of sports analytics.

Rafiq et al. [43] have made significant strides in summarizing sports videos, particularly within cricket. They utilized transfer learning, using the AlexNet scene classification capabilities, to improve scene categorization in cricket videos. By adopting a pre-trained model on a diverse dataset to focus specifically on cricket scenes, they effectively classified video segments such as batting, bowling, and crowd reactions. Their approach demonstrated AlexNet's superiority over other models, like Inception V3 and VGGNet16, by achieving a notable increase in accuracy, up to 99. 26% on smaller datasets. This work improves the utility of sports broadcasts by enabling more focused summaries and offering methodologies that could be adapted to other sports.

Li et al. [33] explored the application of neutrosophy theory to analyze and visualize sports news data. This theoretical framework, which studies the origin, nature, and scope of neutralities, was used to analyze data from significant sporting events using Excel data statistics, Newtonian analysis, and messy dynamics. Their work offers a novel approach to understanding the dynamics of sports media coverage, revealing patterns and evolutions across different media event types. By classifying the development stages of sports media events into the beginning, high-tide, and decay periods, they provided insights into the cyclic nature of media attention and its impact on public engagement with sports.

Fenil et al. [12] introduced a real-time violence detection system in football stadiums, employing Histogram of Oriented Gradients (HOG) and Bidirectional Long Short-Term Memory (BDLSTM) networks. This system processes extensive real-time video feeds to detect vio-

lent actions by analyzing movement patterns and predicting potential escalations. BDLSTM networks leverage past context and future actions to make accurate predictions, enhancing security and spectator safety during live sports events.

Pavitt et al. [40] discussed the use of natural language processing (NLP) and conversational interfaces (CI) to support match analysis and scouting. They highlighted how these AI techniques could provide analytical support to sports professionals, enabling them to quickly delve deeper into traditional data sources. Using NLP and CI, users can interact with complex datasets and analysis outputs through simple conversational interfaces, making advanced data analytics more accessible to a broader range of elite and grassroots sports users.

Ye et al. [55] developed ShuttleSpace, a virtual reality platform to assist badminton coaches and analysts in evaluating players' trajectory data. By allowing coaches to visualize and interact with 3D trajectory data in an immersive environment, they provided a powerful tool for enhancing training sessions and improving player performance analysis. This application of VR in sports training represents a significant leap forward in how coaches can utilize data to refine athletes' techniques and strategies.

Cao et al. [5] created the Sports Gene DataBase (SGDB), an innovative resource that compiles gene expression datasets related to physical activity. By allowing researchers to search for genes expressed before and after exercise and analyze variations across gender, age, and type of exercise, the SGDB offers valuable insights into the genetic impacts of physical activity. This tool could revolutionize personalized sports medicine by enabling more targeted health and performance interventions based on genetic markers.

2.3 Deep Learning in Golf Sports

The detection and analysis of golf clubs using deep learning technologies have become a cornerstone in sports analytics, significantly enhancing the precision of identifying golf clubs in video and image data. This review delves into the evolution and application of convolutional neural networks (CNNs) and other sophisticated models that have facilitated advanced and accurate analytics during golf swing analysis.

CNNs have had a substantial impact on image recognition tasks, extending to the detection of sports equipment. The foundational work by Karpathy et al. [26] utilized various methods to integrate temporal information within sequences of images, setting the groundwork for dynamic object detection, such as golf clubs during swings. This approach laid the foundation for subsequent advancements in sports analytics.

Following this, Simonyan et al. [48] incorporated optical flow into the action recognition architecture, significantly enhancing motion pattern analysis specific to golf clubs. Optical flow allows for more granular movement analysis, distinguishing subtle differences in how other clubs are used during a swing.

LSTM networks have been pivotal in capturing long-term dependencies within sequences, essential for continuously tracking golf clubs during a swing. The integration of LSTMs within a long-term recurrent convolutional network (LRCN) by Donahue et al. [10] has enhanced the model's capability to remember and process features over extended periods. This capability is handy in analyzing entire golf swing sequences, from the initial stance to the follow-through.

Yeung and Russakovsky [56] developed an end-to-end algorithm combining feature extraction and reinforcement learning techniques. This enabled the precise identification of critical moments in golf club movements, focusing on the most informative frames for swing analysis. Such targeted analysis is essential for identifying critical improvements in swing technique.

The synergy between 3D neural networks and object detection models, such as Faster R-CNN, has significantly enhanced spatial recognition capabilities. This integration is particularly effective for accurately locating golf clubs against the complex backgrounds often found in golf footage [17]. Accurate detection is crucial for automated systems used in real-time sports broadcasts and training feedback systems.
The architecture developed by Carreira et al.[6] combines a 3D neural network with a two-stream architecture, significantly enhancing the detection and analysis of complex actions like golf swings. This robust recognition and temporal analysis model is ideal for high-performance analytics in professional sports.

McNally et al. [36] introduced a specialized video database for golf swings, providing a rich training resource for deep-learning models tailored to golf club detection. This database supports the development of targeted and efficient training regimes, enhancing the predictive accuracy of neural networks.

Furthering the analysis of golf swings, Ko and Pan [29] utilized a Bidirectional LSTM network to perform a detailed 3D analysis of swing and body motion using data from a single frontal camera and a motion capture suit. This thorough analysis allows for a precise examination of body sway and head movement, critical for fine-tuning athlete performance.

Gehrig et al. [13] applied single-frame analysis to robustly fit a golf club's location to a swing trajectory model. This approach demonstrates the effectiveness of simpler machine learning algorithms and basic CNNs for quick and accurate swing event classification, which is suitable for real-time applications.

Deep learning has revolutionized the ability of automated systems to detect and analyze golf clubs in sports analytics. From basic CNNs to complex integrations involving LSTM and 3D neural networks, these technologies have significantly improved understanding and performance analysis in golf. Ongoing research and technological advancements are expected to enhance the accuracy and efficiency of these detection systems, offering advanced tools for athletes and coaches.

Chapter 3

Background

In this chapter, we delve into the foundational technologies and advanced neural network architectures that underpin the methodologies applied in this thesis. Our exploration is structured around several key components of modern deep learning critical for processing complex visual data and extracting meaningful information from high-speed video recordings of golf swings.

We begin with a discussion on convolutional neural networks (CNNs), renowned for their ability to efficiently process structured array data such as images and video frames. CNNs form the backbone of many image analysis applications, and their robust feature extraction capabilities make them ideal for the initial stages of interpreting visual data.

Next, we focus on the convolutional block attention module (CBAM), an advanced feature refinement mechanism that selectively emphasizes the most informative features within the data and suppresses less useful ones. By integrating CBAM into CNN architectures, we enhance the network's ability to focus on relevant aspects of the input data, thereby improving the precision of the analysis.

Additionally, we explore recurrent neural networks (RNNs), specifically emphasizing long-short-term memory (LSTM) networks. LSTMs are adept at handling sequences and temporal data, making them particularly useful for analyzing video frames that capture golf swings' dynamic and temporal progression. The ability of LSTM networks to remember long-term dependencies allows for a nuanced understanding of motion sequences, which is crucial for segmenting and classifying the different phases of a golf swing.

We examine advanced neural network architectures such as ResNet50, MobileNetV2, and MobileNetV3 to augment our analytical framework further. These architectures introduce significant innovations in network design, such as residual learning and efficient computation, which are vital for processing high-resolution video data in real time. Each model brings unique strengths to golf swing analysis, from deep feature learning to real-time processing capabilities on mobile devices.

This chapter provides a comprehensive overview of each technology, detailing its architectures, functionalities, and roles in the context of this research. Through this exploration, we establish a solid foundation for applying and discussing these technologies to analyze and improve golf swing techniques.

3.1 Convolution Neural Network (CNN)

Convolution neural networks (CNNs) [38] are specialized deep learning architectures renowned for their proficiency in processing structured array data such as images. These networks automatically detect intricate patterns and features without requiring manual feature extraction, making them ideal for tasks like image classification, object detection, and more. Figure 3.1 demonstrates the basic neural network architecture. Below, we delve into the essential layers of a CNN, exploring their functions, operations, and significance in depth.

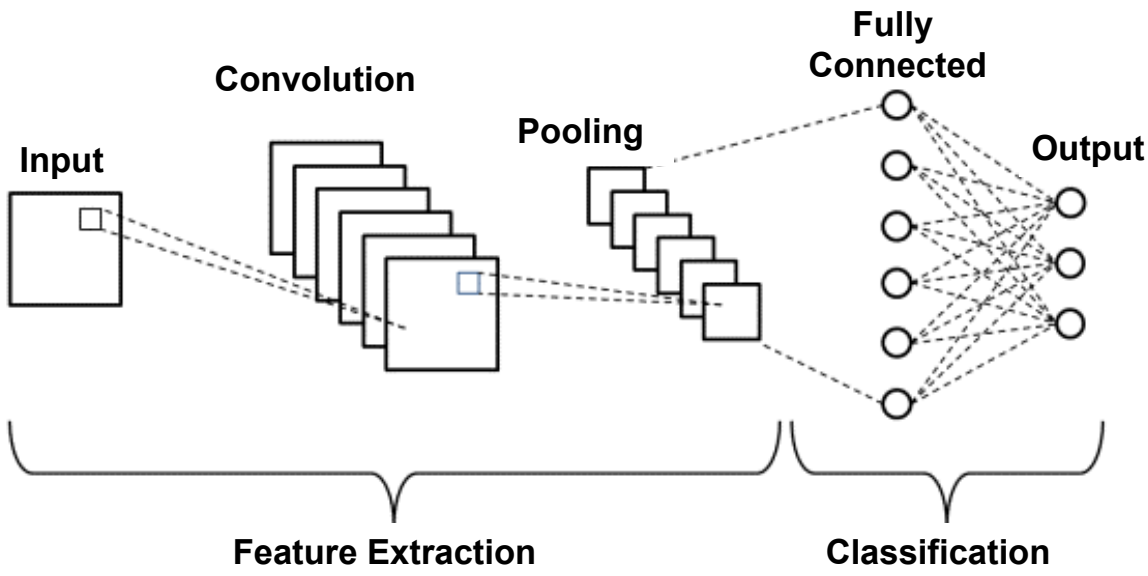

Figure 3.1: Basic convolution neural networks architecture.

3.1.1 Input Layer

The input layer is the gateway into the neural network. For image processing tasks, this layer receives the raw pixel data of the image, formatted as a three-dimensional array where the dimensions correspond to image height H, image width W, and depth D (which represents color channels, typically three for RGB images). The input layer passes these pixel values to the following layers without modification or computation.

3.1.2 Convolution Layer

The convolution layer is the fundamental building block of a CNN. It consists of a set of learnable filters (or kernels), which are small spatially (height and width) but extend through the full depth of the input volume. As these filters slide, or convolve, across the image, they perform element-wise multiplications and produce feature maps. These feature maps represent spatial hierarchies of features in the image, such as edges in the initial layer, followed by textures and patterns in deeper layers.

Mathematically, the convolution operation involves computing the dot product between the entries of the filter and the input, which can be represented as:

$$a_{ij}^{l} = \sigma\left(b^{l} + \sum_{m}\sum_{p=0}^{P_l-1}\sum_{q=0}^{Q_l-1} w_{pq}^{lm} a_{i+p,j+q}^{l-1}\right)$$

where σ denotes a nonlinear activation function, b^l represents the bias term, w_{pq}^{lm} is the weight of the m-th filter at position (p, q), and P_l, Q_l are the dimensions of the filter. This layer's parameters are shared among all spatial locations, significantly reducing the number of parameters and computational cost.

3.1.3 ReLU Layer

Following the convolution layer, the Rectified Linear Unit (ReLU) layer applies a non-linear activation function. The ReLU function is defined as $f(x) = \max(0, x)$, and it introduces non-linearity to the system, allowing the network to solve more complex problems. The ReLU function is preferred over other activation functions like sigmoid or tanh due to its computational efficiency and the ability to mitigate the vanishing gradient problem.

3.1.4 Pooling Layer

Pooling (or subsampling or downsampling) layers follow ReLU operations to reduce the spatial size of the representation, decrease the number of parameters and computations in the network, and control overfitting. The most common forms are max pooling and average pooling:

$$a_{ij}^l = \max\left\{a_{mn}^{l-1} | m \in [i \cdot s, i \cdot s + F], n \in [j \cdot s, j \cdot s + F]\right\}$$

While average pooling returns the average of all values from the same region, max pooling returns the maximum value from the image area the kernel covers. s represents the stride of the convolution operation, and F is the filter size.

3.1.5 Fully Connected Layer

Towards the end of the network, fully connected layers connect every input from the previous layer to every output neuron. These layers flatten the high-level features learned by prior convolution layers and combine all features across the image. The output from this layer is computed as:

$$a^l = \sigma\left(W^l a^{l-1} + b^l\right)$$

where W^l and b^l denote the weight matrix and bias vector, respectively, and σ is an activation function, typically a softmax or sigmoid for classification tasks.

3.1.6 Output Layer

The final layer, typically the output layer, uses the softmax function for multi-class classification tasks, where it converts logits, the numerical outputs of the last linear layer of a multi-class classification neural network, into probabilities by comparing them with other class logits. The softmax function is:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

for each class i, where z_i are the logits for class i. This function ensures the output values are distributed between 0 and 1, summing up to 1, making them interpretable as probabilities.

3.2 Convolutional Block Attention Module (CBAM)

The convolutional block attention module (CBAM) [52] is an attention mechanism for neural networks that improves feature representation by focusing on the most vital features and suppressing the less relevant ones. This module is applied after a convolutional layer before

passing the feature maps to subsequent layers or modules. It operates sequentially through two distinct components: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM), as shown in Figure 3.2.

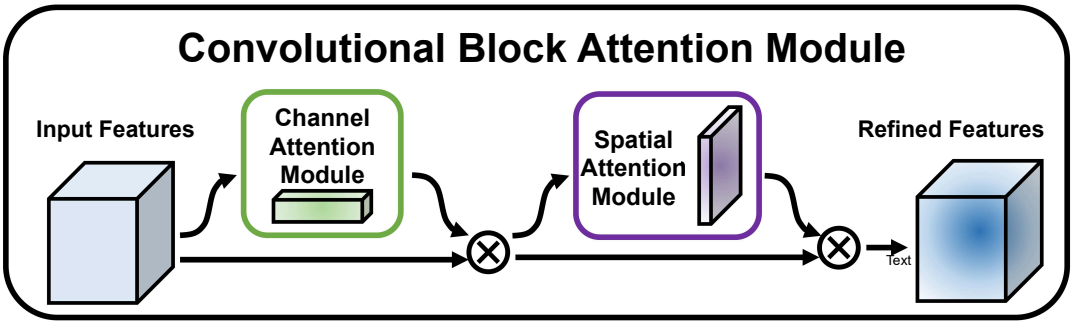

Figure 3.2: Overview of convolutional block attention module.

3.2.1 Overview

CBAM is designed to be a lightweight and general module that can be seamlessly integrated into any convolutional neural network architecture. It sequentially infers attention maps along two separate dimensions, channel and spatial, thereby adapting the feature maps. This process enhances features vital for the specific task. It suppresses background noise or irrelevant details, improving performance in visual tasks like image classification, object detection, and segmentation.

3.2.2 Channel Attention Module (CAM)

The Channel Attention Module (CAM) focuses on what is meaningful along the channel dimension. It exploits the inter-channel relationship of features by utilizing the global spatial information of feature maps, thus enabling the network to emphasize informative features while suppressing less useful ones.

Mathematical Formulation

Given an input feature map $F \in \mathbb{R}^{C \times H \times W}$, where C is the number of channels, and H and W are the height and width of the feature map, respectively, CAM first applies global average pooling and global max pooling across spatial dimensions to generate two different spatial context descriptors:

$$F^c_{avg} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{cij},$$

$$F^c_{max} = \max_{i \leq H, j \leq W} F_{cij},$$

where F_{cij} denotes the (i, j)-th spatial element of the c-th channel.
These descriptors are then forwarded through a shared network, typically a multi-layer perceptron (MLP) with one hidden layer, to produce the channel attention map. The MLP is applied separately to each descriptor, and their outputs are summed and passed through a sigmoid function to generate the final channel attention map $M_c \in \mathbb{R}^{C \times 1 \times 1}$:

$$M_c = \sigma\left(MLP(F^c_{avg}) + MLP(F^c_{max})\right),$$

where σ represents the sigmoid activation function, ensuring that the attention weights are normalized between 0 and 1.

3.2.3 Spatial Attention Module (SAM)

After channel-wise attention has refined the features, the Spatial Attention Module (SAM) focuses on where an essential spatial location is in the feature map. SAM highlights salient features that are spatially important for the task at hand.

Mathematical Formulation

The input to SAM is the feature map re-weighted by the channel attention. SAM applies average-pooling and max-pooling along the channel axis to highlight informative regions, generating two complementary spatial context descriptors:

$$F^s_{avg} = \frac{1}{C}\sum_{k=1}^{C} F_{kij},$$

$$F^s_{max} = \max_{k \leq C} F_{kij},$$

where F_{kij} denotes the (i, j)-th spatial element of the k-th channel.
These two maps are then concatenated and convolved with a standard convolution layer to generate the spatial attention map $M_s \in \mathbb{R}^{1\times H\times W}$:

$$M_s = \sigma\left(f^{7\times 7}([F^s_{avg}; F^s_{max}])\right),$$

where $f^{7\times 7}$ denotes a convolution operation with a 7×7 filter, and σ is the sigmoid function. This map is used to re-weight the spatial regions of the feature map, focusing the network's attention on important spatial locations.

3.2.4 Integration into CNNs

CBAM is modular and can be integrated within a CNN architecture at various points. It is typically placed after each convolutional block before activation functions, allowing the network to adaptively adjust the weighting of features through both channel and spatial dimensions. The effectiveness of CBAM in enhancing feature representations leads to significant improvements in CNNs' performance across various visual tasks.

3.3 Recurrent Neural Networks and Long Short-Term Memory

Recurrent Neural Networks (RNNs) are a class of neural networks that excel at processing sequential data by maintaining a 'memory' of previous inputs using their internal state. They are ideally suited for time series prediction, speech recognition, language modeling, and other tasks where the temporal sequence of data is crucial. Long-short-term memory (LSTM) networks, a special kind of RNN, are designed to overcome the limitations of traditional RNNs, particularly in learning long-range dependencies.

3.3.1 Recurrent Neural Networks (RNNs)

RNNs process an input sequence one element at a time, maintaining in their hidden layers a state that implicitly contains information about the history of all past elements of the sequence. This allows them to exhibit dynamic temporal behavior for a time sequence. Unlike feedforward neural networks, RNNs have a loop within them, allowing information to persist, as shown in Figure 3.3.

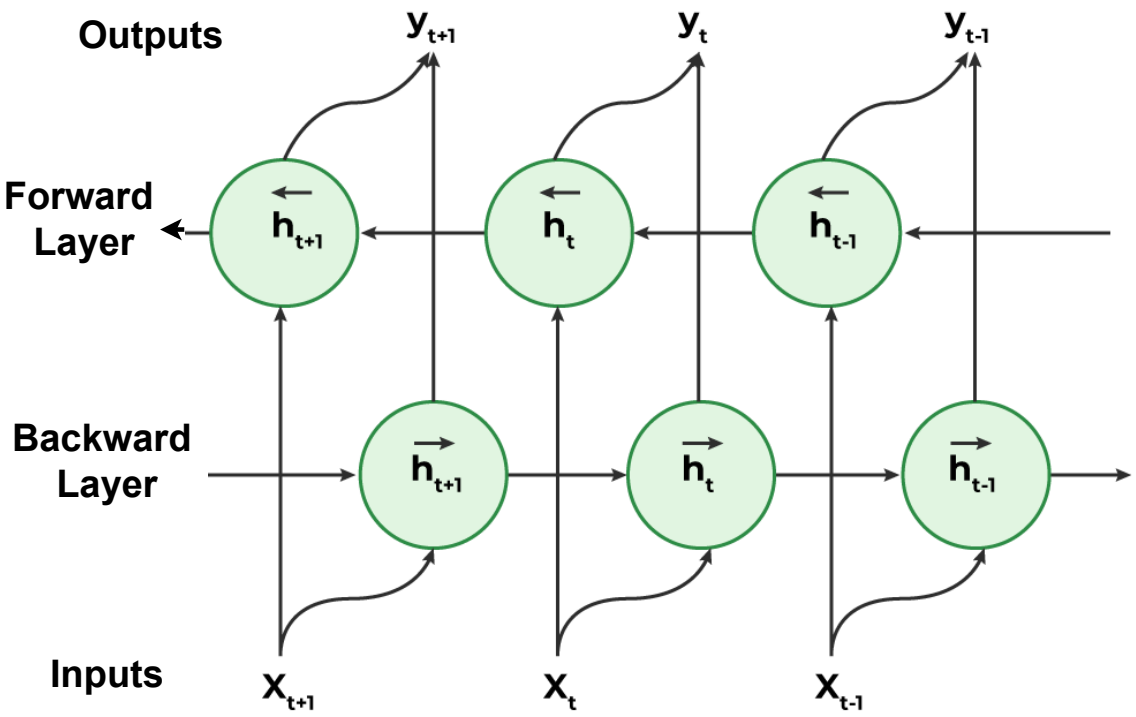

Figure 3.3: Overview of Bi-directional Recurrent Neural Network.

Mathematical Formulation of RNN

The basic model of an RNN uses the current input and the previous hidden state to compute the current hidden state. The hidden state at any time step t is given by:

$$h_t = \sigma(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

where x_t is the input at time step t, h_t is the hidden state at time step t (also the 'memory' of the network), W_{hh} is the weight matrix for connections between hidden units of adjacent time steps, W_{xh} is the weight matrix for connections between input and hidden units, b_h is the bias, and σ is the activation function, typically a non-linear function like tanh or ReLU.

Challenges with RNNs

While theoretically robust, RNNs in practice suffer from significant training difficulties due to problems like vanishing and exploding gradients. These problems arise during backpropagation through time (BPTT) when gradients propagated over many time steps tend to vanish (become very small) or explode (become very large), making long-term dependencies hard to learn.

3.3.2 Long Short-Term Memory (LSTM) Networks

LSTM [20] networks are an advanced type of RNN designed specifically to avoid the long-term dependency problem. They do this by introducing 'gates' that regulate the flow of information. These gates can learn which data in a sequence is important to keep or throw away.

Architecture of LSTM

An LSTM unit typically contains three types of gates:

- **Forget Gate:** Decides what information should be discarded from the cell state. It looks at h_{t-1} and x_t, and outputs a number between 0 and 1 for each number in the cell state C_{t-1}.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

- **Input Gate:** Decides what new information is added to the cell state.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

- **Output Gate:** Determines what the next hidden state should be, which is used in the output of the LSTM unit.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \cdot \tanh(C_t)$$

These gates allow the LSTM to let information through selectively, based on the strength and relevance of the input data at each step.

Cell State Update

The cell state, which is the key to LSTMs, is updated using the following equations:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

where C_t is the new cell state, f_t is the output from the forget gate, i_t is the output from the input gate, and $\tilde{C}_t$ is the new candidate values, scaled by how much we decided to update each state value.

3.3.3 Integration into CNNs with CBAM

In scenarios where CBAM enriches features extracted from CNNs, passing these features into an LSTM allows the model to capture temporal dependencies and dynamics effectively. This combination is particularly potent in applications like video analysis, where understanding the evolution of scenes or actions is crucial.

3.4 Advanced Neural Network Architectures

This section delves into the innovative deep-learning architectures of ResNet50, MobileNetV2, and MobileNetV3. These models represent transformative advancements in neural network design and have been adapted in this research to significantly enhance the accuracy and efficiency of extracting features from high-speed video data. This is crucial for analyzing the rapid and intricate motions observed in golf swings.

3.4.1 ResNet50

ResNet50 [19], a variant of the Residual Network with 50 layers, is a cornerstone in the evolution of convolutional neural networks. It is primarily known for its ability to solve the vanishing gradient problem that plagues traditional deep networks as they become more profound.

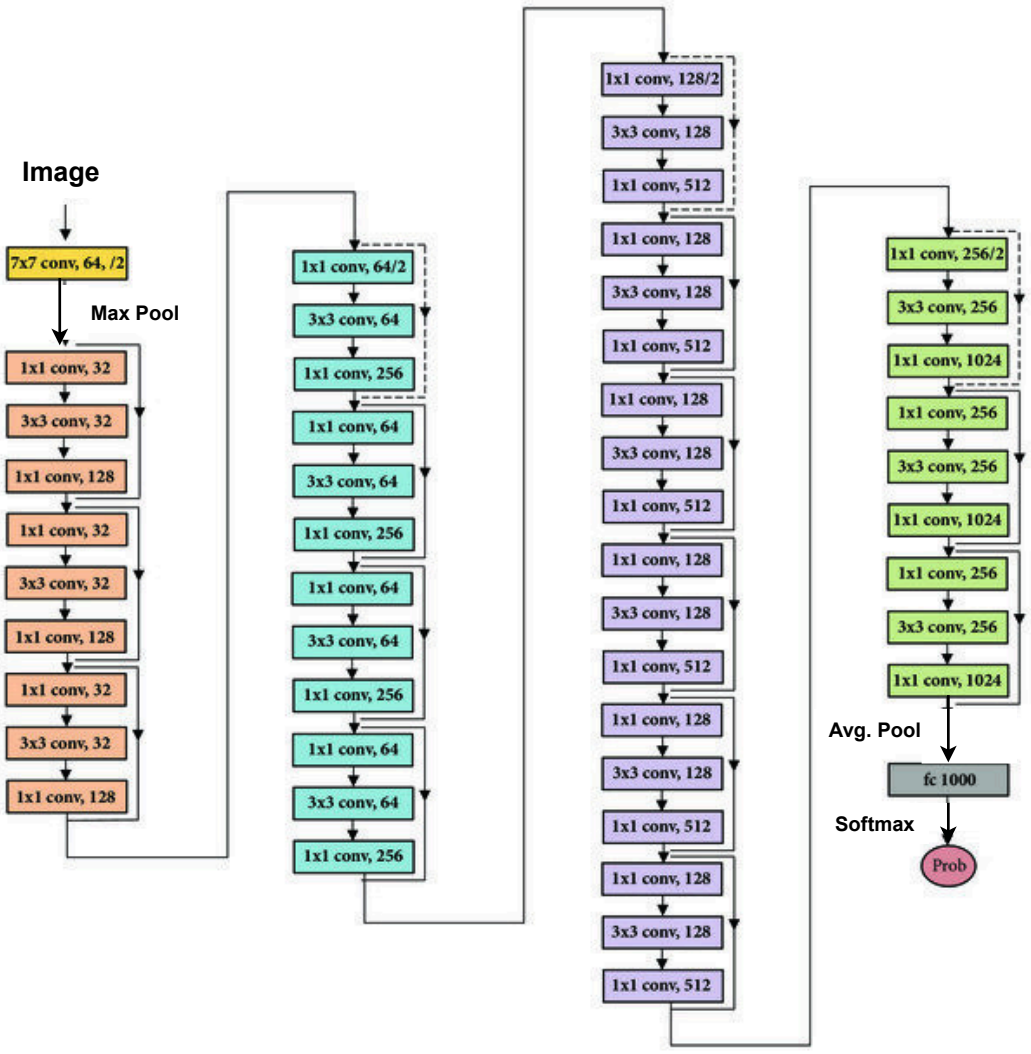

Figure 3.4: Block diagram of Resnet-50

Architecture and Functionality

The core innovation in ResNet50 is its use of residual blocks, which incorporate skip connections, allowing gradients to bypass one or more layers during backpropagation. This architectural innovation is critical in maintaining a strong gradient flow across many layers, facilitating the training of much deeper networks than was previously feasible.

$$R(x) = F(x, \{W_i\}) + x$$

Here, x represents the input to the layer, $F(x, \{W_i\})$ represents the residual function to be learned, and $R(x)$ is the output. The skip connections, by simplifying the path for the gradients, mitigate the issue of vanishing gradients, thereby preserving the learning capability of the network even with increased depth.

Applications in Golf Swing Analysis

ResNet50's ability to function effectively in deep learning models makes it exceptionally suitable for detailed image analysis, such as that required in high-resolution video data of golf swings. It can be trained to identify distinct phases within the golf swing by recognizing

subtle patterns and features, providing a detailed analysis that is invaluable for coaching and performance enhancement.

3.4.2 MobileNetV2

Following the introduction of MobileNet, which aimed to deliver CNN capabilities to mobile devices efficiently, MobileNetV2 [46] enhances this design by optimizing the balance between latency and accuracy. It introduces an architecture tailored for high efficiency with minimal performance loss, making it suitable for real-time applications.

Architecture and Efficiency

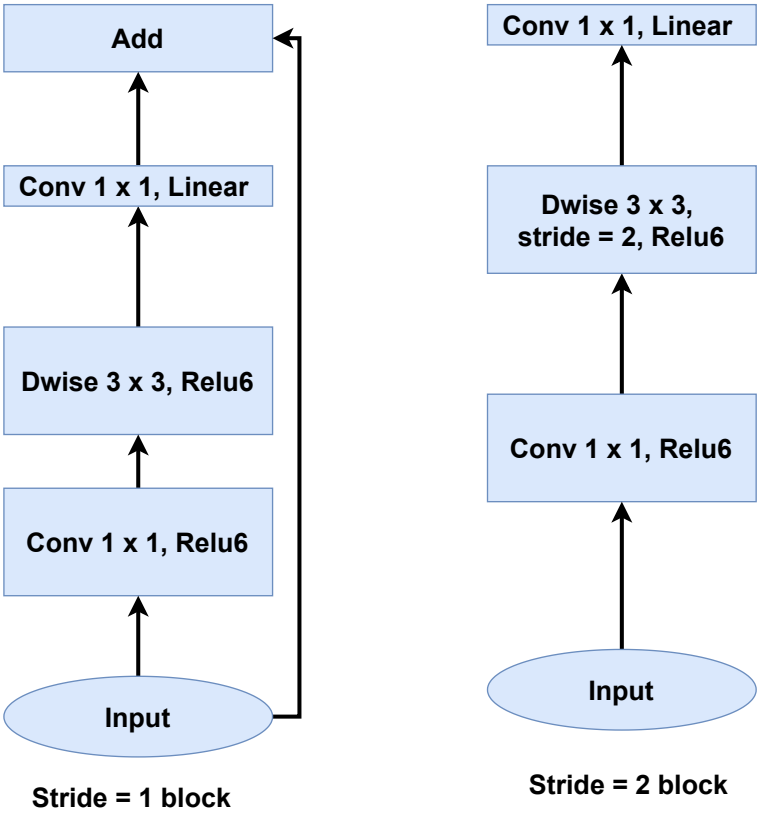

Figure 3.5: Overview of MobileNetV2 Convolutional Blocks

MobileNetV2 introduces innovative concepts such as linear bottlenecks and inverted residuals. Linear bottlenecks control the flow of channels, effectively reducing the dimensionality of the data within the network, thus preserving computational resources. The inverted residuals incorporate lightweight depthwise separable convolutions that significantly reduce the model's size and complexity without compromising its ability to process complex features.

$$F(x) = \text{conv\_bn}(\text{relu6}(\text{conv\_bn}(\text{relu6}(\text{conv\_bn}(x)))))$$

This structural design enhances MobileNetV2's efficiency, making it highly applicable for mobile applications that require real-time analysis, such as dynamic sports analytics, where immediate computational response is crucial.

Utilization in Golf Swing Analysis

MobileNetV2's swift and efficient computation provides near-real-time video processing capabilities, which are crucial for live sports performance analysis. It enables the detailed

tracking and recognition of golf swing phases, facilitating immediate feedback and insights essential for training and improvement.

3.4.3 MobileNetV3

MobileNetV3 [22] represents the latest advancement in the MobileNet series of architectures. It was developed to provide high-accuracy results with minimal computational resources, making it ideal for mobile and edge devices. This architecture has been fine-tuned through automated machine-learning techniques that optimize its layers and operations based on the constraints and capabilities of the hardware it is intended to run on.

Innovations in Architecture

MobileNetV3 synthesizes two distinct approaches in neural network design: the hardware-aware Neural Architecture Search (NAS) and the NetAdapt algorithm. NAS focuses on automating the design of neural network architectures by searching through a predefined space of potential designs to find the most efficient architecture for a given task and hardware setup. This approach combines insights from the NetAdapt algorithm, which iteratively trims a pre-trained model to improve efficiency while maintaining or enhancing performance.

The architecture of MobileNetV3 is characterized by its segmented design, which incorporates the best practices from its predecessors and new technologies developed through state-of-the-art research. Key features include:

Squeeze-and-Excitation (SE) blocks: These blocks adaptively recalibrate channel-wise feature responses by explicitly modeling channel interdependencies. They significantly enhance the representational power of the network by focusing on useful features and suppressing less useful ones. Implementing SE blocks in MobileNetV3 is optimized for mobile efficiency, reducing computational overhead while boosting performance.

Modified Depthwise Separable Convolutions: These convolutions, which separate the filtering and combining convolution steps, are central to the MobileNet architecture. In MobileNetV3, they are further optimized to reduce latency and parameter count.

H-Swish Activation Function: MobileNetV3 introduces the h-swish activation function, an approximation of the Swish function optimized for hardware efficiency. This function is employed in the non-linear transformations within the network, providing a balance between computational efficiency and non-linear representational power.

$$L(x) = x \cdot \text{relu6}(x + 3)/6$$

Learning Feature Extraction

MobileNetV3'sIts sophisticated layer designs and activation functions enhance the capacity for learning effective feature extraction. The network learns to focus on the input data's most significant features, optimizing the architecture's processing paths to minimize redundancy and maximize informational throughput. This targeted feature extraction is particularly beneficial for tasks requiring high accuracy levels from video input, such as dynamic sports analytics.

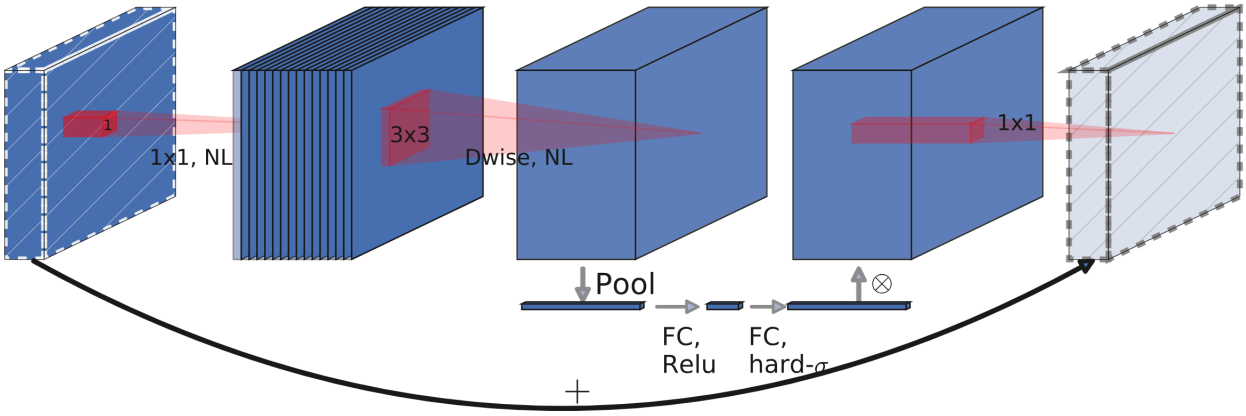

Figure 3.6: MobileNetV3 (MobileNetV2 + Squeeze-and-Excite)

Integrating SE blocks allows MobileNetV3 to perform context-aware feature enhancement, dynamically adjusting the emphasis on certain features based on the input data. This capability ensures that the network remains responsive to the nuances of complex visual patterns, such as those involved in the precise movements of a golf swing.

Applications in Golf Swing Analysis

In the context of golf swing analysis, MobileNetV3's advanced capabilities enable real-time, on-device processing of video data, which is critical for providing immediate feedback. Its efficient computation allows for faster analysis and conserves energy, which is crucial for mobile applications. The real-time processing capabilities of MobileNetV3 ensure that every minute adjustment in a player's swing is captured and analyzed, providing highly accurate and actionable feedback. This immediate analysis helps make quick corrections, leading to improved performance and technique refinement over time.

Chapter 4

GolfDB Dataset

This section will discuss the publicly available GolfDB [36] dataset, a carefully selected extensive video data set for use in golf recognition, and, more specifically, to pioneer the task of sequencing golf swings. GolfDB encompasses a substantial collection of more than 390,000 movie frames depicting 1400 distinct golf swings. Regarding domain-specific datasets, GolfDB truly stands out as a significant resource. To our knowledge, GolfDB represents the first large dataset explicitly designed for computer vision applications in golf.

GolfDB was compiled by assembling 580 YouTube videos featuring golf shots in both real-time and slow motion. For the task of golf swing sequencing, maintaining continuous visibility of the club shaft throughout the swing is of paramount importance. To ensure the absence of visual disturbances caused by motion blur, the dataset exclusively considered high-quality videos for inclusion in GolfDB. This collection of YouTube videos predominantly showcases seasoned players hailing from renowned tours such as the PGA, LPGA, and Champions Tours, encompassing a diverse pool of 248 individuals with distinct golf swing styles.

These videos were sourced from various perspectives and locations on golf courses, including the driving range, tee boxes, fairways, and sand traps. The dataset's strength lies in its deliberate inclusion of a wide array of variables, including different players, clubs, perspectives, lighting conditions, surroundings, and native frame rates, all of which enhance its capacity to facilitate the generalization of computer vision models. It's worth noting that the YouTube videos in GolfDB were sampled at a standard 30 frames per second and possess a resolution of 720p.

4.1 Golf Swing Events

In [15], soccer events were settled in one second, while golf swing events can be contained in a single frame due to the strict definition they are. There have been many ideas about different golf swing events [30], but in this research, we have considered that a golf swing sequence is made up of eight events that happen one after the other, and each has its definition:

- Takeaway (A): This happens right before the takeaway, at the frame, before the backswing makes any visible movement.
- Toe-up (TU): This is when the length of the golf club shaft is level with the ground during the backswing.
- Mid-backswing (MB): The golfer's arm is straight out from their body during the backswing.

- Bottom (T): The bottom event is when the golf club changes direction, going from the backswing to the downswing.
- Mid-downswing (MD): The golfer's arm is straight out of his body during the downswing.
- Impact (I): This event marks when the golf clubhead hits the ball.
- Mid-follow-through (MFT): At this point in the follow-through phase, the head of the golf club is level with the ground.
- Finish (F): The finish event happens right before the player strikes their last pose, which marks the end of the swing.

Figure 4.1: Swing Events from one of the subjects from the GolfDB dataset.

Together, these eight events make up a complete golf swing routine. Each one happens in a particular order and corresponds to a different part of the swing.

4.2 Golf Swing Sequencing

The golf swing sequencing challenge involves identifying critical times within edited movies by capturing individual golf swings [15]. There are various reasons for the decision to focus

on trimmed golf swing videos.

The main objective of the research is to offer immediate biomechanical feedback in real-world settings, as opposed to the limited scope of analyzing golf swings solely within broadcast videos. Although GolfDB possesses the necessary data for spatial and temporal localization of entire golf swings in untrimmed movies, it warrants further investigation.

Furthermore, anyone involved in golf, including golfers and golf instructors, can analyze a golf swing sequence in real-time by conveniently capturing a controlled video of a singular golf swing using a mobile device. This process guarantees that the subject of interest is correctly positioned within the frame. This prevents the necessity for intricate spatial and temporal localization.

Finally, a video sample that includes a singular instance of a golf swing consists of a distinct series of events that transpire in a particular and predetermined arrangement. The utilization of sequential information can potentially improve the precision of event detection.

4.3 Annotation

Using an internal MATLAB algorithm, 1400 trimmed video samples of golf swings from the YouTube video collection were retrieved, replicating the methods employed by the original authors. Subsequently, this code was distributed to four annotators to reproduce the annotation procedure.

The annotators were tasked with identifying complete golf swings in each YouTube video while excluding chip shots, putts, and pitch shots. Each sample was required to pinpoint specific frames, including the beginning, eight pivotal moments in the golf swing, and the finish. It's important to note that the number of frames between the start of the sample and the "Address" event and between the "Finish" event and the end of the sample naturally varied. In addition, some samples included practice swings at the beginning of the study. Due to the samples' inherent frame rate, pinpoint accuracy in event labeling was impossible. For instance, in real-time samples shot at their native frame rate of 30 frames per second, capturing the exact moment of impact was a rarity. In such cases, the annotators were instructed to exercise discretion and select the frame closest to the event's occurrence.

In addition to labeling events, the annotators drew bounding boxes around the clubhead and the golf ball throughout the swing. They were also required to specify the type of club and view and indicate whether the sample was in slow motion or real-time, with slow motion defined as 30 frames per second. The annotators extracted player names from the video titles and determined the gender of each participant by cross-referencing the names with available online resources. The annotators received domain-specific information to ensure accurate annotation before commencing the procedure. Subsequently, a skilled golfer conducted a rigorous quality verification process on the dataset, following the methodology of the original authors.

Chapter 5

Methodology

5.1 Data Prepocessing

Data preprocessing is a critical step in ensuring the quality and utility of data before it is used in machine learning models. In the thesis, the preprocessing steps are designed to handle and transform video data from YouTube videos that capture golf swings, which are essential for the subsequent deep-learning tasks. The main goal of the preprocessing steps is to extract relevant frames from each video, resize them to uniform dimensions, and standardize the input for efficient learning.

The data originates from a .mat file containing structured data extracted from golf-related videos. This data includes multiple attributes such as video IDs, player information, event markers, and bounding boxes for each swing event. The first step involves loading this MATLAB file, which is then converted to a structured dictionary. Each key-value pair in the dictionary corresponds to different attributes of the dataset. This dictionary is then transformed into a Pandas DataFrame for easier manipulation. The columns are named appropriately to represent each attribute, such as "id," "youtubeid," "player," and "events," among others.

The next step is to clean and normalize the data by ensuring that all entries are uniform. This includes converting multidimensional arrays into singular values where necessary and ensuring identifiers like YouTube IDs and player information are stored in a consistent format. This is crucial for accurately indexing and accessing specific videos during the frame extraction phase.

Using the cleaned DataFrame, each video file associated with a specific YouTubeid is processed to extract frames corresponding to the golf swing events marked in the events column. The frames are extracted using OpenCV, and each frame is cropped according to the bounding box (bbox) coordinates provided for that specific swing. The extracted frames are then resized to a fixed dimension to ensure uniformity, which is critical for the consistency of CNN models. Padding is added to maintain aspect ratio integrity using the mean pixel values from the ImageNet dataset. This helps normalize the data against a common background used in pre-trained models.

After processing, the videos are saved in a structured directory that allows easy access during the training phase. Each video is stored in a format that preserves the video ID and the selected dimension, making it straightforward to retrieve during model training. Finally, the dataset is split into training and validation sets based on predefined splits in the dataset. This is essential for evaluating the model's performance independently from its training, ensuring that the trained model generalizes well to new unseen data. Each split is saved separately, preserving the integrity of the testing framework.

5.2 Data Loader

The data loader is an essential component of the machine learning pipeline, particularly effective in managing video and image data for training deep learning models. This section elaborates on the data loader's role in efficiently managing video data, extracting frames, and preparing them for model input, optimizing memory usage and computational efficiency.

The data loader initiates by loading a structured data set from a specified file containing preprocessed video clips and associated metadata, such as event labels and identifiers. Each video, representing a sequence of actions in a golf swing, is accompanied by multiple frames. The directory in which these videos are stored is predefined, and the sequence length to be sampled from each video is specified.

During the training phase, the data loader randomly selects a starting frame within each video to capture variability and augment the training data. Then it samples several consecutive frames to ensure that the model learns from various points within the golf swing sequence. In contrast, for testing, the loader processes each frame sequentially from the video to provide a comprehensive evaluation across the entire swing.

Each frame is labeled according to its relevance to specific golf swing events. Frames corresponding to key swing events are tagged with specific labels. In contrast, those not directly aligned with such events receive a default label, indicating their noncritical nature in swing dynamics.

Upon loading, frames undergo several preprocessing steps to ensure they are in the appropriate format for the neural network:

- **Color Adjustment:** Frames are converted from BGR to RGB color format to align with the standard color channel order expected by most pre-trained deep learning models.
- **Transformation and Normalization:** The data loader applies transformations, such as converting image data arrays into tensors suitable for processing by PyTorch models. It also normalizes the images using predefined mean and standard deviation values, which helps stabilize the training process by ensuring the input features are on a similar scale.

The data loader is implemented using a multithreaded approach where multiple workers load data in parallel to enhance data handling efficiency, especially given the large size of video data. This reduces the time the model waits for data, accelerating the training process.

The loader batches the data, grouping multiple samples into a single batch. This is crucial for training deep learning models, as batch processing helps in gradient estimation while optimizing the neural network, providing a more stable and reliable convergence during training.

The data loader is seamlessly integrated with PyTorch's ecosystem, utilizing its `Dataset` and `DataLoader` classes. This integration facilitates the customization of data handling processes, such as shuffling the data for training (to prevent the model from learning spurious patterns) and ensuring that batches are dropped if they are not complete, according to the training requirements.

The data loader is vital to the deep learning pipeline, ensuring that data is efficiently processed, transformed, and ready for the neural network. By handling the complexities of data manipulation, the data loader allows the model to focus solely on learning from the data,

thereby enhancing the overall effectiveness and efficiency of the machine-learning workflow. This consistent approach to data handling ensures that the data is always optimally presented for learning, regardless of the experimental setup or the specific model used.

5.3 Proposed Architecture

The architecture proposed in this study is motivated by the significant advancements and success of deep learning models in human activity recognition (HAR) and sports analytics, specifically in detecting and analyzing complex sequences like golf swings. The use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs), particularly long short-term memory (LSTM) networks, has proven effective in HAR due to their ability to handle spatial and temporal data, respectively. CNNs excel in extracting robust spatial features from video frames, as demonstrated by Ronao and Cho [45], who showed superior performance of CNNs over traditional methods in activity recognition. Additionally, hybrid models that combine CNNs and RNNs, like those discussed by Lv et al. [35], have significantly improved the recognition of continuous and dynamic activities by capturing both spatial and temporal dependencies.

Recent advancements in attention mechanisms, such as the Convolutional Block Attention Module (CBAM) introduced by Woo et al. [52], have further enhanced the performance of deep learning models in HAR. CBAM emphasizes the most informative features by applying channel and spatial attention sequentially, which helps the network focus on relevant parts of the input data. This mechanism is crucial for accurately detecting and analyzing complex activities, where distinguishing between subtle movement differences is necessary. The use of transfer learning, where models pre-trained on large datasets are fine-tuned on domain-specific data, has shown great promise in improving the performance of HAR systems. Ha and Choi [18] demonstrated that transfer learning could significantly enhance model accuracy and efficiency by leveraging pre-existing knowledge from large-scale datasets. This approach is particularly useful in sports analytics, where collecting large, labeled datasets can be challenging.

In golf swing analysis, detecting and classifying different phases of a golf swing requires capturing fine-grained details from video sequences. The application of CNNs for spatial feature extraction and LSTMs for temporal sequence modeling has proven effective. McNally et al. [36] created a specialized video database for golf swings, highlighting the importance of tailored datasets in training accurate models. Additionally, Ko and Pan [29] utilized bidirectional LSTMs for detailed 3D analysis of golf swings, emphasizing the need for comprehensive temporal modeling to capture the complexity of the motion.

The proposed architecture delineates an advanced, state-of-the-art neural network design to detect and analyze events within golf swing sequences. The neural network encompasses three core modules: the feature extraction module, the feature refinement module, and the temporal module. This design strategy aims to capture golf swing videos' intricate temporal and spatial characteristics. The overarching goal is to extract detailed feature representations, enhance these features via attention mechanisms, and then analyze their temporal evolution to predict the type and timing of golf swing events.

The first stage of the architecture is the Feature Extraction Module, which leverages a pre-trained MobileNetV3 model. MobileNetV3 is chosen for its balance between performance and computational efficiency, making it ideal for extracting high-quality spatial features from video frames. As each video frame passes through the MobileNetV3 network, rich feature

representations are generated that capture the essential spatial characteristics necessary for subsequent analysis.

Next, the extracted features are passed through the Feature Refinement Module, which incorporates the Convolutional Block Attention Module (CBAM). CBAM sequentially applies channel and spatial attention to the features, effectively highlighting the most informative parts of the input data while suppressing irrelevant information. This refinement process ensures that the network focuses on the critical aspects of the golf swing, such as the golf club's position and movement, enhancing the model's overall discriminative power.

Finally, the refined features are fed into the Temporal Module, which consists of LSTM networks. LSTMs are adept at capturing long-term dependencies and modeling temporal sequences, making them well-suited for analyzing the evolution of golf swing events over time. By processing the sequence of refined features, the LSTM networks can learn the temporal patterns and transitions between different phases of the golf swing, enabling accurate detection and classification of events within the swing sequence.

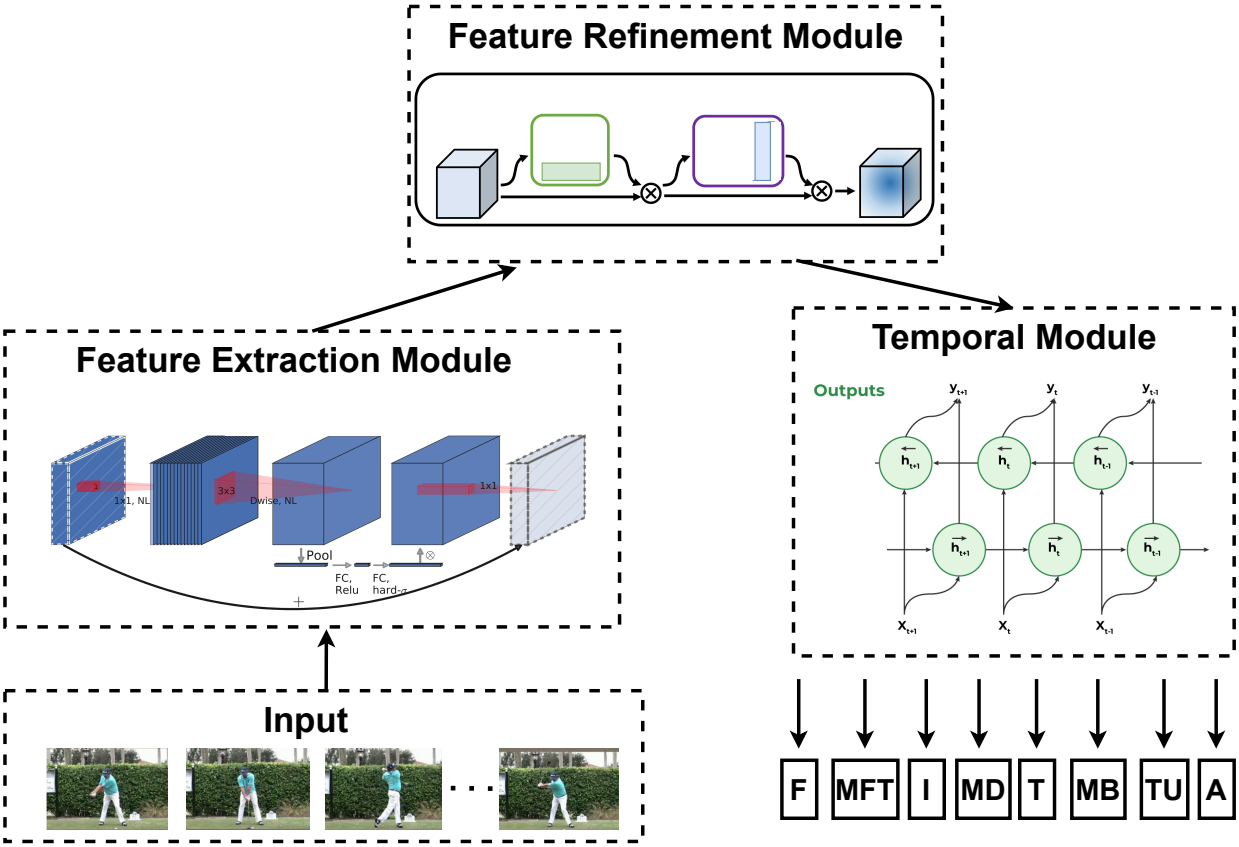

Figure 5.1: Schematic representation of the proposed event detector neural network architecture. The architecture integrates a MobileNetV3-based Feature Extraction Module with Channel and Spatial Attention mechanisms for feature refinement, followed by a Bidirectional LSTM Temporal Module for capturing dynamic temporal dependencies.

The data flow through the proposed architecture begins with individual video frames being processed by the Feature Extraction Module, which extracts spatial features using MobileNetV3. These features are then refined by the Feature Refinement Module, where CBAM enhances the most relevant aspects of the data. The refined features are subsequently passed to the Temporal Module, where LSTM networks analyze the temporal dynamics of the sequence. This sequential processing ensures that the golf swing's spatial and temporal characteristics are effectively captured and analyzed, leading to precise predictions of golf swing

events.

By integrating these modules, the proposed architecture effectively combines the strengths of CNNs, attention mechanisms, and LSTMs to create a powerful tool for golf swing analysis. This comprehensive approach ensures that the model can handle the task's complexity, providing accurate and reliable results that can significantly enhance sports performance analytics.

5.3.1 Input

The input layer is the input layer at the inception of the event detector's processing pipeline. It receives a meticulously curated sequence of video frames, each a snapshot of a golf swing at a particular instant. The preprocessing pipeline is responsible for standardizing the video data, involving:

- Resizing frames to ensure a uniform resolution amenable to consistent processing by the neural network.
- Normalizing pixel values across the dataset to the same scale to facilitate effective feature extraction.
- Optionally augmenting the dataset through rotation, translation, and scaling techniques to bolster the model's robustness against overfitting and improve its generalization capabilities.

5.3.2 Feature Extraction Module

The Feature Extraction Module functions as the cerebral cortex of the Event Detector, transforming each frame into a rich tapestry of features. The chosen backbone for this module is the state-of-the-art MobileNetV3 architecture. With its pedigree in balancing efficiency with high-performance feature extraction, it is an optimal choice for real-time applications where computational resources are at a premium.

MobileNetV3 Architecture

MobileNetV3 represents a confluence of innovations from automated machine learning (AutoML) and insights into efficient neural architecture designs tailored for hardware constraints. This model employs depthwise separable convolutions, a breakthrough technique significantly reducing computational complexity without sacrificing feature extraction capability. Leveraging a pre-trained MobileNetV3 allows the model to capitalize on knowledge distilled from the expansive ImageNet dataset, providing a diverse and rich feature initialization.

Model Adaptation

Adaptation of MobileNetV3 within the Event Detector framework involves repurposing the architecture from its original image classification role. The classifier layer designed for ImageNet's 1000-class problem is replaced with an output that unfurls a multidimensional feature map. This map captures an abstract representation of the input frame's content. An additional non-linear activation layer then processes these feature maps, enhancing the network's ability to model complex, non-linear relationships within the visual data.

5.3.3 Feature Refinement Module

The Feature Refinement Module sharpens the network's focus. Employing attention mechanisms enhances the expressiveness of the features procured from the extraction phase.

Channel Attention

Channel Attention is a discerning filter that evaluates and amplifies the most informative feature channels. At its core lies an adaptive pooling layer that condenses each channel to a single representative value. A compact, fully connected network then examines these values to compute a scaling factor for each channel, governed by a sigmoid activation that ensures these factors are within the range (0, 1). The output of this mechanism is a channel-wise weighted feature map tailored to emphasize salient features crucial for recognizing the intricate phases of a golf swing.

Spatial Attention

In concert with Channel Attention, Spatial Attention emphasizes the 'where'—the significant locations in the feature map. The module generates a spatial attention map by aggregating features across channels using average and max pooling and applying a dedicated convolutional layer. Once processed through a sigmoid function for normalization, this map enhances the original feature map, ensuring the network focuses on regions rich with discriminatory information.

5.3.4 Temporal Module

The refined features serve as fodder for the Temporal Module, which deciphers the sequential nature of a golf swing. By leveraging LSTM networks, the module weaves a coherent narrative of the frames, unraveling the temporal patterns that characterize the golf swing.

Long Short-Term Memory Network

The LSTM network stands out for its proficiency in mitigating the infamous vanishing and exploding gradient problems that plague standard RNNs. Its intricate design, featuring memory cells and meticulously engineered gating mechanisms, empowers the network with an enduring memory. This memory carries relevant historical information throughout the sequence, permitting the network to make informed predictions based on a comprehensive temporal context.

Bidirectional LSTM

Implementing a bidirectional LSTM enriches the model's temporal acuity by allowing it to look backward and forward in time. This duality enables each frame to be understood in isolation and as part of a continuum, echoing the holistic approach a human expert might take in analyzing a golf swing.

Regularization and Output Layer

Regularization, instantiated through dropout, serves as the network's safeguard against the specter of overfitting. The network is coerced into developing more robust and generalized features by stochastically turning off a subset of neurons during training. The LSTM's temporally informed output is funneled through a fully connected linear classifier, which delineates the probabilities of each frame belonging to various golf swing segments.

5.3.5 Output

The culmination of the event detector's pipeline is a sequence of predictions, each mirroring a frame from the original video input. The model's outputs portray a probabilistic landscape over the different segments of a golf swing, including the backswing, downswing, impact,

and follow-through. Analyzing this sequence unveils the swing's granular, frame-by-frame temporal profile, yielding invaluable coaching and performance enhancement insights.
With its assembly of designed modules, the Event Detector provides advancements in deep learning architectures. It processes raw video data through a gauntlet of convolutional and recurrent networks; each module is fine-tuned to tease out and interpret the complex spatial-temporal interplay of a golf swing. The architecture's end-to-end design, culminating in a sequence of event predictions, offers a potent analytical tool to dissect and enhance the mechanics of a golf swing.

5.4 Model Summary

Table 5.1 presents a concise summary of the proposed model architecture, detailing the type of each layer, the output shape, the number of parameters, and the connections between layers.

Table 5.1: Detailed Summary of the EventDetector Model Architecture

Layer Type	**Output Shape**	**Kernel Size**	**Stride**	**Padding**	**Parameters**
Conv2dNormActivation	(3, 224, 224)	(3, 3)	(2, 2)	(1, 1)	432
InvertedResidual	(16, 112, 112)	(3, 3)	(1, 1)	(1, 1)	-
InvertedResidual	(24, 56, 56)	(1, 1)	(1, 1)	(0, 0)	-
InvertedResidual	(24, 56, 56)	(3, 3)	(2, 2)	(1, 1)	-
InvertedResidual	(40, 28, 28)	(1, 1)	(1, 1)	(0, 0)	-
InvertedResidual	(40, 28, 28)	(3, 3)	(1, 1)	(1, 1)	-
InvertedResidual	(80, 14, 14)	(1, 1)	(1, 1)	(0, 0)	-
InvertedResidual	(80, 14, 14)	(3, 3)	(2, 2)	(1, 1)	-
InvertedResidual	(80, 14, 14)	(1, 1)	(1, 1)	(0, 0)	-
InvertedResidual	(112, 14, 14)	(1, 1)	(1, 1)	(0, 0)	-
InvertedResidual	(112, 14, 14)	(3, 3)	(1, 1)	(1, 1)	-
InvertedResidual	(160, 7, 7)	(1, 1)	(1, 1)	(0, 0)	-
InvertedResidual	(160, 7, 7)	(3, 3)	(1, 1)	(1, 1)	-
InvertedResidual	(160, 7, 7)	(1, 1)	(1, 1)	(0, 0)	-
InvertedResidual	(160, 7, 7)	(3, 3)	(1, 1)	(1, 1)	-
Conv2dNormActivation	(960, 7, 7)	(1, 1)	(1, 1)	(0, 0)	153600
ChannelAttention	(960, 7, 7)	-	-	-	11520
SpatialAttention	(960, 7, 7)	(7, 7)	(1, 1)	(3, 3)	-
LSTM	(None, 512)	-	-	-	1315840
Dropout	(None, 512)	-	-	-	-
Linear	(None, 9)	-	-	-	4617

5.5 Model Training

The training regimen for the Improved Event Detector is designed to fine-tune the neural network's parameters systematically. This process equips the model with the ability to precisely detect and temporally analyze complex events depicted within sequences of golf swings.

5.5.1 Training Configuration

A carefully structured training configuration is the foundation of the model's learning process. This setup dictates the model's interaction with the dataset through multiple iterations, known as epochs, where each epoch represents a complete pass through the data. Setting

the number of epochs to 50 strikes a balance between ensuring sufficient exposure to the data for adequate learning and maintaining computational efficiency.

Checkpoint is an essential strategy employed after each epoch. It involves saving the model's current state, thus documenting its progression. These checkpoints are critical as they allow for evaluating the model's performance over time and provide a mechanism for resuming training from specific points, thereby enhancing the flexibility and depth of longitudinal performance assessments.

5.5.2 Computational Resource Allocation

Efficient training of the model necessitates the strategic allocation of computational resources. This is achieved by employing parallel processing techniques, which are essential for expediting the training data throughput and thus accelerating the training process.

The allocation of processing units is optimized based on the available system resources to prevent potential bottlenecks in data handling. This optimization ensures a continuous and smooth supply of data for processing, which is crucial for maintaining the momentum of the training process.

5.5.3 Batch Formulation

Batch processing plays a pivotal role in the learning algorithm. The batches, made up of sets of frames that play one after the other, are designed to make good use of memory and keep the learning process's randomness, which is essential for gradient descent to work.

The organization of data into batches that reflect the natural sequence of events within a golf swing is critical for training the model to recognize and interpret these temporal patterns accurately. This sequential data presentation helps the model understand the progression and dynamics of golf swings.

5.5.4 Model Architecture Initialization

The initialization of the model architecture marks a fundamental phase in the training process, wherein the network's structure is configured to effectively capture and interpret the temporal dynamics present in the sequential data of golf swings. The Improved Event Detector model incorporates several advanced architectural features for analyzing these sequences' time-dependent aspects.

The model's ability to understand temporal dependencies lies in its use of bidirectional Long Short-Term Memory (LSTM) layers. These layers are designed to process data through the sequence in both forward and backward directions. This dual-path processing allows the model to have insights into future and past contexts simultaneously, which is critical for capturing the complete temporal dynamics of a golf swing. The LSTM layers are configured with 256 hidden units each, providing a robust framework to learn complex patterns and dependencies in the data.

The initialization step also sets the stage for adaptive learning, configuring the model to adjust its internal parameters dynamically. This adaptability is key to refining the model's performance in response to the complexities of the input data encountered during training.

5.5.5 Data Standardization

Before entering the training phase, the input data is subjected to a rigorous standardization process. This normalization ensures the input features are consistent and uniformly scaled, facilitating the model's ability to process and learn from the data effectively.

The scaling parameters for normalization are derived from the ImageNet dataset, which serves as a benchmark for feature scaling. This alignment ensures that the model's inputs are standardized according to well-established norms, promoting consistency in model training.

5.5.6 Optimization Framework

The cross-entropy loss function guides the optimization of the model during training. This function evaluates the model's predictions against the actual data, providing a quantitative performance measure. Additionally, to accommodate the variance in class distribution within the data, the loss function is modified with weights to balance the influence of each class on the model's learning.

The Adam optimizer, known for its efficiency in handling sparse gradients and adaptive learning rates, is employed to optimize the model's parameters. This choice supports a nuanced adjustment of parameters, enhancing the convergence rate and overall effectiveness of the training process.

The model's performance is continuously monitored throughout training through metrics such as loss and accuracy. These metrics provide insights into the effectiveness of the training interventions and the model's evolving capability to classify events accurately.

Tools like the AverageMeter utility aggregate these metrics, offering a running tally of performance that helps track and analyze the model's progress throughout the training period.

5.5.7 State Preservation and Learning Progression

A systematic checkpointing mechanism that preserves the model's state at each epoch adds to the iterative learning process. This preservation is crucial for maintaining a record of incremental advancements and ensuring that each step in the model's evolution is captured.

The backpropagation algorithm plays a central role in learning and adjusting the model's parameters based on the calculated gradients. This mechanism is fundamental to refining the model's accuracy and enhancing its predictive capabilities.

5.5.8 Training Iterations

The training loop forms the core of the learning process. The model actively engages in forward and backward passes in this loop, interspersed with parameter updates. This iterative cycle is vital for continuously improving the model's performance.

The structured approach to training the Improved Event Detector ensures a comprehensive enhancement of the model's capabilities. By adhering to rigorous deep learning principles, the model is fine-tuned to perform detailed analyses of golf swing events with remarkable precision and reliability.

5.6 Model Evaluation

The evaluation process for the CBAM-based Improved Event Detector aims to assess its effectiveness in accurately detecting and temporally analyzing golf swing events depicted within video sequences. This section outlines the methodologies, metrics, and considerations employed to gauge the model's performance.

5.6.1 Evaluation Configuration

A meticulously structured evaluation configuration is pivotal in thoroughly assessing the model's capabilities. This setup dictates the interaction between the model and the evaluation dataset, facilitating comprehensive analysis and performance measurement.

The choice of dataset for evaluation is a critical determinant of the evaluation's validity and relevance. The dataset should encompass diverse golf swing scenarios representative of real-world scenarios to provide a comprehensive testbed for the model.

The sequence length used during evaluation is carefully selected, considering the temporal dynamics inherent in golf swing events. This parameter effectively influences the model's ability to capture and analyze temporal dependencies.

5.6.2 Computational Resource Allocation

Efficient evaluation of the model requires strategic allocation of computational resources, ensuring timely processing of evaluation data and expedited performance analysis.

Similar to training, parallel processing techniques are employed during evaluation to maximize computational efficiency. These techniques help minimize processing time while maintaining accuracy in performance assessment.

5.6.3 Evaluation Metrics

The evaluation metrics are quantitative measures to assess the model's performance and efficacy in accurately detecting and classifying golf swing events.

Phase-wise accuracy measures the model's ability to correctly identify each phase of the golf swing, providing insights into its temporal analysis capabilities. This metric is calculated by comparing the model's predictions with ground truth annotations for each phase.

The model's accuracy reflects its overall performance in correctly classifying golf swing events across all phases. This metric provides a holistic view of the model's efficacy in event detection and classification.

5.6.4 Performance Assessment

The performance assessment process involves analyzing the model's predictions against ground truth annotations and identifying strengths, weaknesses, and areas for improvement.

The confusion matrix offers a detailed breakdown of the model's predictions across different event classes, highlighting any misclassification patterns or confusion.

Error analysis involves scrutinizing instances of misclassification or erroneous predictions to discern underlying causes and potential areas for model refinement.

5.6.5 Model Generalization

The evaluation process also assesses the model's generalization capabilities, gauging its performance on unseen data and its ability to extrapolate learned patterns to new scenarios.

Cross-validation techniques are employed to assess the model's performance across multiple folds of the evaluation dataset, ensuring robustness and generalizability of results.

The model's potential for transfer learning is evaluated by assessing its performance on datasets or scenarios different from those used during training. This analysis provides insights into the model's adaptability to diverse contexts.

Chapter 6

Results and Discussions

This chapter comprehensively analyzes the experiments conducted to explore the effectiveness of various deep-learning models in sequencing golf club swings. The primary goal of these experiments is to identify the most effective configuration of neural network architectures and enhancements for analyzing video data from the GolfDB dataset. Achieving high precision in swing phase detection is crucial for coaching purposes and enhancing overall golf performance.

The experimental sequence is designed to build progressively, starting from simpler models and advancing towards more complex and refined architectures. Initially, the experiments begin with MobileNetV2 paired with LSTM to establish a performance baseline. The study then progresses to include more sophisticated architectures like ResNet50 coupled with LSTM and MobileNetV3 paired with LSTM, each chosen for their unique architectural strengths and capabilities in handling complex video data. The culmination of our research is the innovative model that integrates MobileNetV3 with the Convolutional Block Attention Module (CBAM) and LSTM, aiming to harness the best features of each component to improve the accuracy and efficiency of swing phase detection significantly.

6.1 MobileNetV2 + LSTM

The initial experiment employs MobileNetV2 in conjunction with LSTM to set a baseline for golf swing phase detection. This model configuration was selected for its proven computational efficiency and effective feature extraction capabilities from video frames, while LSTM is utilized to model the temporal relationships between sequential frames.

6.1.1 Experimental Setup

This setup involved pre-training MobileNetV2 on the ImageNet dataset to leverage its robust feature extraction capabilities. It was then fine-tuned on the GolfDB dataset, which contains richly annotated video data of various golf swings. The LSTM layers were appended to capture the dynamic temporal progression evident across the golf swing sequence, focusing on detecting subtle nuances between different swing phases.

6.1.2 Implementation Details

The training was executed on an NVIDIA Tesla GPU with a batch size of 32, using the Adam optimizer with an initial learning rate of 0.001. The objective was to minimize the loss between the predicted swing phases and the ground truth, effectively optimizing the synergy between MobileNetV2's feature extraction capabilities and LSTM's sequential data handling.

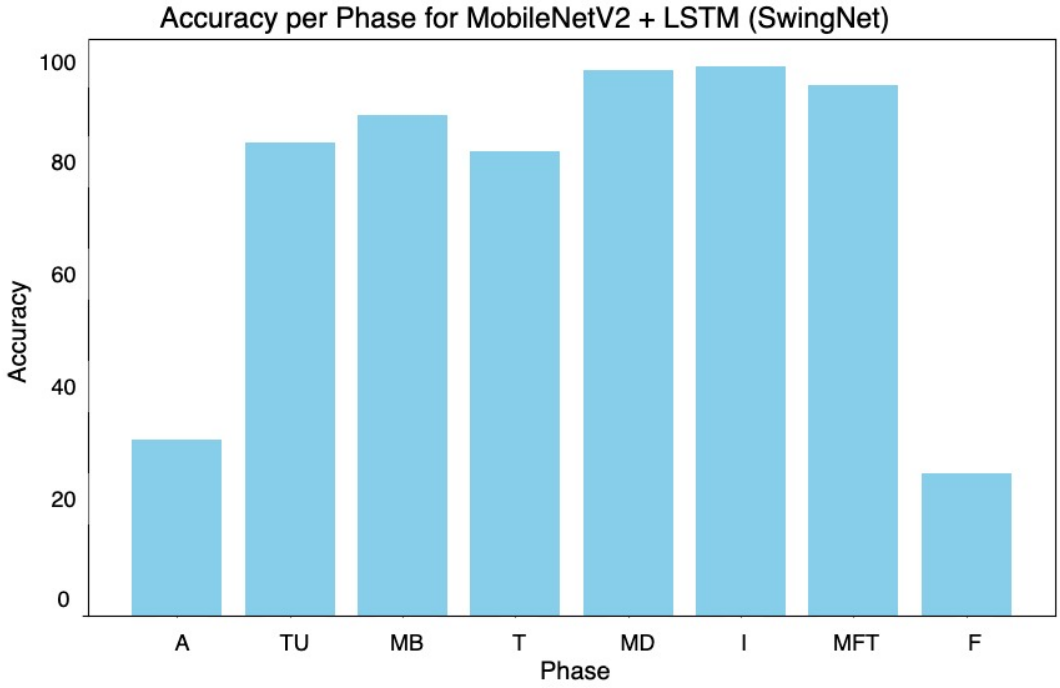

Figure 6.1: Accuracy per Phase for Swing Net (MobileNetv2 + LSTM)

6.1.3 Results

The results from this first set of experiments, which utilized MobileNetV2 coupled with LSTM to detect various phases in a golf swing, are graphically represented below.
The data reveals:

- Exceptionally high accuracy in the **Impact (I)** phase, nearing 100%, underscoring the model's effectiveness in detecting the swing's most dynamic and visually distinctive part.
- Significant accuracies in the **Mid-Downswing (MD)** and **Mid-Follow-Through (MFT)** phases, indicating the model's proficiency in capturing clear, significant motions.
- Lower accuracies in the **Address (A)** and **Finish (F)** phases, potentially due to the static nature of these moments where less distinct motion occurs, posing challenges for the LSTM component in detecting relevant temporal features.

6.1.4 Discussion

The variance in detection accuracies across different swing phases can be attributed to several factors:

Dynamic vs. Static Phases: Phases such as **Impact** and **Mid-Follow-Through**, which involve rapid and significant changes, are more effectively tracked by the LSTM due to their clear temporal discrepancies. In contrast, static phases like **Address** and **Finish** are characterized by minimal movement, which challenges the LSTM's ability to detect significant temporal features due to the lack of dynamic content.

Subjective Labeling Variability: The phases with lower accuracies often need more subjective judgments in their labeling within the dataset, introducing inconsistencies that can

confuse the model. This variability reduces the model's effectiveness in accurately classifying these less dynamic phases.

Temporal Localization Challenges: The accurate temporal localization of static events within a video sequence is inherently challenging. These events typically do not exhibit robust distinguishing features, which can lead to lower detection accuracies.
To enhance the detection accuracy in static phases, several strategies could be considered:

- **Enhanced Feature Extraction:** Incorporating advanced spatial attention mechanisms like CBAM could improve the model's focus on subtle spatial cues within static frames, potentially enhancing recognition accuracy.
- **Data Augmentation:** Implementing targeted data augmentation strategies for static phases could increase the model's exposure to these critical moments, aiding in learning more robust features.
- **Refined Labeling Techniques:** Standardizing the labeling process and incorporating expert reviews could reduce subjectivity and enhance the quality of the training data.

While the MobileNetV2 + LSTM configuration shows promising results for dynamic swing phases, model architecture and data handling enhancements are crucial for improving performance across all phases. This foundational experiment sets the stage for subsequent evaluations involving more complex architectures like ResNet50 and MobileNetV3 with LSTM, leading to the proposed integration of CBAM.

6.2 ResNet50 + LSTM

Following the initial experiments with MobileNetV2 + LSTM, the next configuration tested was ResNet50 paired with LSTM. This model was selected due to ResNet50's more profound architecture, which can capture more complex features from the video data, potentially improving the accuracy of swing phase detection.

6.2.1 Experimental Setup

ResNet50 was pre-trained on the ImageNet dataset to utilize its deep residual learning framework, effectively avoiding the degradation problem that typically occurs with deeper networks. This model was then fine-tuned on the GolfDB dataset to adapt its capabilities to the golf swing phase detection task. The LSTM layers were integrated to maintain the temporal coherence of the swing sequences, aiming to capitalize on the deeper and more nuanced feature extraction provided by ResNet50.

6.2.2 Implementation Details

The combined ResNet50 + LSTM model was trained using the same NVIDIA Tesla GPU as the previous experiments, with a batch size of 32 and an Adam optimizer. The learning rate was initially set to 0.001 but was adjusted based on the validation loss performance during training. The primary objective was to reduce the prediction error across all phases of the golf swing, optimizing the system to handle both the increased depth of the convolutional network and the complexities of temporal sequence modeling.

6.2.3 Results

The accuracies obtained from the ResNet50 + LSTM model are summarized below and discussed subsequently:

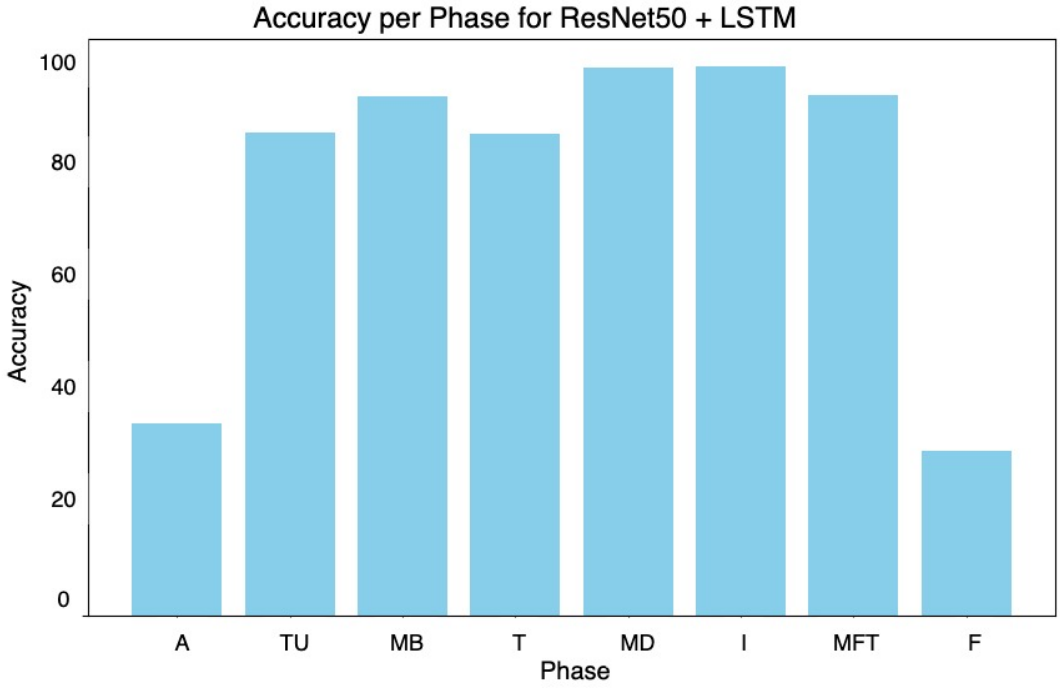

Figure 6.2: Accuracy per Phase for ResNet50 + LSTM

The system's overall accuracy was recorded at 77.55%, indicating a significant improvement from the baseline model. The phase-wise accuracies are as follows:

- The **Impact (I)** phase continued to show very high accuracy, similar to the MobileNetV2 + LSTM model, benefiting from the deep feature extraction capabilities of ResNet50.
- Notable improvements were observed in **Mid-Downswing (MD)** and **Mid-Follow-Through (MFT)**, with accuracies surpassing those from the previous experiment, suggesting that the additional depth provided by ResNet50 enhances detection in dynamically complex phases.
- However, the **Address (A)** and **Finish (F)** phases still showed lower performance, albeit slightly improved from the previous model configurations.

6.2.4 Discussion

The use of ResNet50 with LSTM brings forth several insights:

Enhanced Detection in Dynamic Phases: The deep residual learning of ResNet50 allows for more effective feature extraction during rapid movements, which likely contributes to the higher accuracies in dynamic phases such as Impact and Mid-Follow-Through.

Challenges in Static Phases Persist: Despite improvements, static phases like Address and Finish continue to present challenges. These phases may benefit from further methodological adjustments, such as more sophisticated data augmentation techniques or advanced temporal feature engineering, to better capture the subtleties of static postures.

Overall System Performance: The improvement in overall accuracy to 77.55% underscores the potential of using deeper networks for complex activity recognition tasks like golf

swing analysis. However, the persistent issues in certain phases suggest that simply increasing model depth is not a panacea and must be complemented by other strategies, such as enhanced data preprocessing, improved labeling accuracy, and perhaps the incorporation of attention mechanisms to focus the model on relevant features better.

This experiment sets a robust foundation for further exploration into even more sophisticated models and configurations, leading to the next phase of experiments involving MobileNetV3 with LSTM, followed by the integration of the CBAM.

6.3 MobileNetV3 + LSTM

Building upon the insights gained from previous experiments, the third configuration tested was MobileNetV3 combined with LSTM. MobileNetV3 is known for its efficiency and effectiveness in mobile environments, which comes from lightweight, depthwise separable convolutions and architecture optimized with automated machine learning techniques. This experiment evaluated whether the enhancements in MobileNetV3 could translate into improved detection of golf swing phases, especially in a complex sequence analysis scenario like golf swings.

6.3.1 Experimental Setup

MobileNetV3 was pre-trained on the ImageNet dataset to harness its advanced architecture optimized for performance and efficiency. The model was then fine-tuned on the GolfDB dataset to tailor its capabilities to the golf swing phase detection task. Similar to previous setups, LSTM layers were integrated to analyze the temporal dynamics within the swing sequences, aiming to capitalize on MobileNetv3's enhanced feature extraction capabilities.

6.3.2 Implementation Details

The combined MobileNetV3 + LSTM model was trained on an NVIDIA Tesla GPU, maintaining a consistent batch size of 32 and utilizing the Adam optimizer with an initial learning rate of 0.001. This setup aimed to minimize the loss between the predicted and actual swing phases, focusing on refining the integration of advanced feature extraction with temporal data processing.

6.3.3 Results

The performance results for the MobileNetV3 + LSTM model are illustrated below:
The overall accuracy achieved was 76.975%, with the following phase-wise accuracies observed:

- The **Impact (I)** phase continued to demonstrate very high accuracy, consistent with previous models, reflecting the model's capability to recognize the most dynamic phase of the swing effectively.
- Accuracies in the **Mid-Downswing (MD)** and **Mid-Follow-Through (MFT)** phases were slightly lower compared to the ResNet50 + LSTM setup but still remained robust.
- Similar to earlier experiments, the **Address (A)** and **Finish (F)** phases exhibited lower accuracy, highlighting ongoing challenges with static phases.

6.3.4 Discussion

The implementation of MobileNetV3 with LSTM offers several key takeaways:

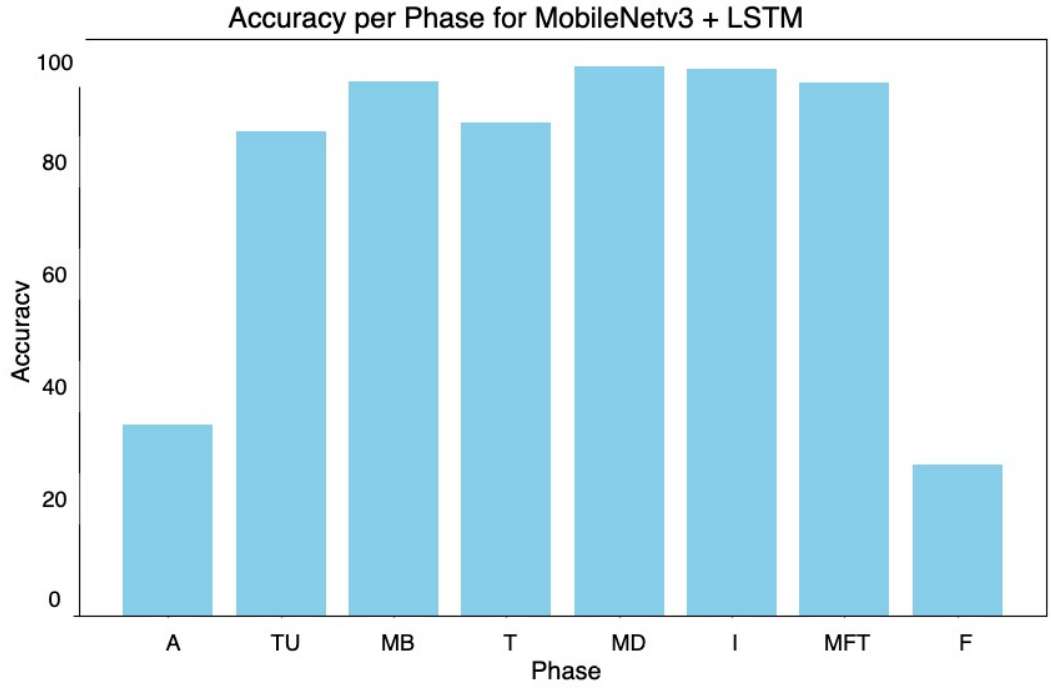

Figure 6.3: Accuracy per Phase for MobileNetV3 + LSTM

Efficient Feature Extraction: MobileNetV3's use of lightweight architectures and attention mechanisms potentially aids in more efficient processing and feature extraction, which is crucial for the rapid dynamics of golf swings.

Performance in Static vs. Dynamic Phases: While dynamic phases such as Impact show consistently high performance due to clear visual and motion cues, static phases like Address and Finish still pose significant challenges. These phases lack pronounced motion changes, making detecting distinct patterns difficult for temporal models.

Comparison to Previous Models: Despite its architectural advantages, MobileNetV3 + LSTM did not significantly outperform ResNet50 + LSTM in overall accuracy, suggesting that factors beyond mere architectural depth and efficiency, such as model tuning and data preprocessing, play critical roles in performance enhancements.

The results from this experiment highlight the nuanced trade-offs between model complexity, efficiency, and accuracy in detecting various phases of the golf swing. The insights gained here will inform the final set of experiments involving the integration of CBAM with MobileNetV3 and LSTM, aimed at addressing the specific challenges identified in detecting less dynamic phases.

6.4 MobileNetV3 + CBAM + LSTM

The culmination of our experimental analysis involves integrating the Convolutional Block Attention Module (CBAM) with MobileNetV3 and LSTM. This configuration was proposed to harness the specific strengths of each component—MobileNetV3's efficient feature extraction, CBAM's focus on relevant features through attention mechanisms, and LSTM's capability to analyze temporal sequences. The objective was to significantly enhance the accuracy and efficiency of detecting golf swing phases, particularly improving the model's performance in static phases.

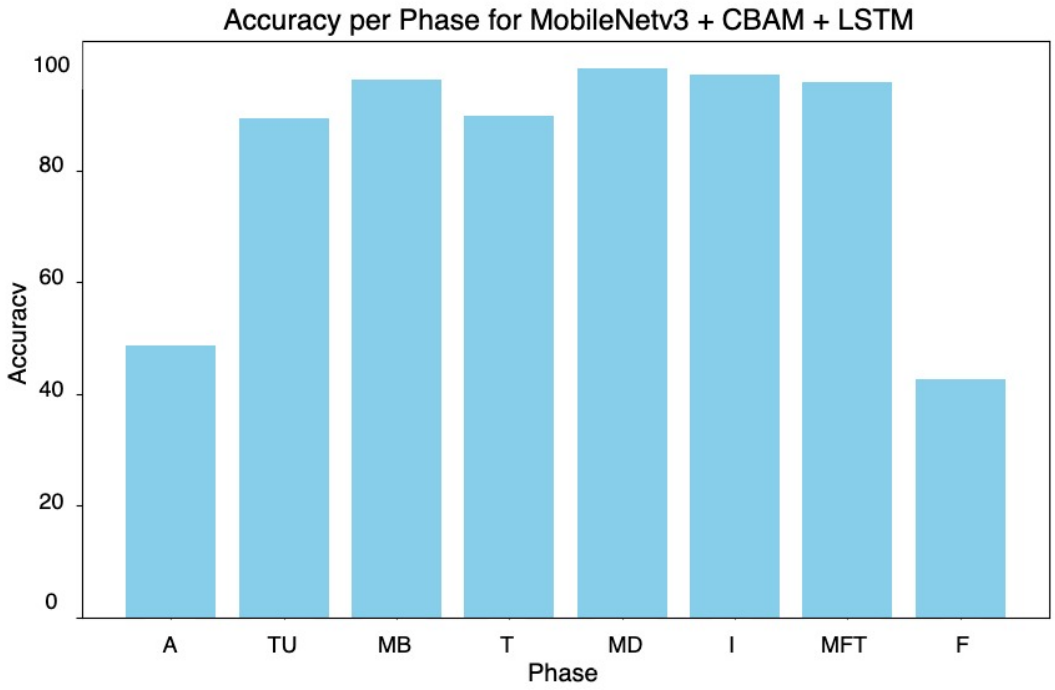

Figure 6.4: Accuracy per Phase for MobileNetV3 + CBAM + LSTM

6.4.1 Experimental Setup

This experiment leverages MobileNetV3's architecture optimized for mobile devices, combined with CBAM to refine the focus on essential features within the video frames and LSTM to handle the temporal dependencies. The model was pre-trained on the ImageNet dataset and subsequently fine-tuned on the GolfDB dataset, similar to previous setups but with CBAM to amplify the model's attention to subtle yet crucial aspects of each swing phase.

6.4.2 Implementation Details

The training was carried out on an NVIDIA Tesla GPU, with parameters adjusted to optimize convergence. The Adam optimizer was used, starting with a learning rate of 0.001, and adjusted dynamically based on the training progress. Including CBAM required adjustments to the training process to ensure that the attention mechanisms were effectively learning to prioritize relevant features.

6.4.3 Results

The performance results for the MobileNetV3 + CBAM + LSTM model are illustrated below, showing the accuracy for each phase of the golf swing:
The overall accuracy achieved was 82.22%, an improvement over all previous configurations. The phase-wise accuracies observed are as follows:

- Consistently high accuracy in the **Impact (I)** phase, similar to earlier experiments.
- Notable improvement in **Mid-Downswing (MD)** and **Mid-Follow-Through (MFT)** phases, reflecting the beneficial effects of integrating CBAM.
- Significant enhancements in the **Address (A)** and **Finish (F)** phases, where earlier models struggled. These phases now show markedly better accuracy, demonstrating CBAM's effectiveness in highlighting less obvious features in these more static swing parts.

6.4.4 Discussion

The integration of CBAM with MobileNetV3 and LSTM offers a promising solution to previously identified challenges:

Enhanced Focus on Relevant Features: CBAM's ability to prioritize spatial and channel-wise features significantly contributes to the model's improved performance, particularly in phases involving subtle movements or less dynamically expressive movements.

Balanced Attention Across Phases: The attention mechanisms help balance the model's sensitivity across different phases, ensuring that dynamic and static phases are analyzed accurately.

Model Complexity and Efficiency: While the addition of CBAM increases the model's complexity, the benefits in terms of detection accuracy justify the additional computational overhead. This configuration effectively balances depth and efficiency, achieving superior performance without excessively burdening computational resources.

This experiment demonstrates the effectiveness of CBAM when integrated with MobileNetV3 and LSTM and sets a new standard for the accuracy of golf swing phase detection.

6.5 Discussion

This study extensively evaluated the performance of four deep learning models in detecting various phases of golf swings as part of ongoing efforts to enhance coaching tools and analytical techniques in sports. The models tested were MobileNetV2 + LSTM, ResNet50 + LSTM, MobileNetV3 + LSTM, and MobileNetV3 + CBAM + LSTM, each selected for their distinct architectural features and potential utility in handling complex video data.
The initial approach in this study involved using MobileNetV2 coupled with LSTM, which was selected for its renowned computational efficiency and efficacy in processing dynamic activities. MobileNetV2, specifically designed for high-performance scenarios on mobile devices, leverages a lightweight architecture that facilitates rapid processing and lower computational costs. When combined with LSTM, which excels in capturing temporal dynamics across sequences, this configuration forms a robust baseline ideal for real-time applications where quick and efficient processing of time-series data is critical. This pairing aims to harness the strengths of MobileNetV2 in handling pronounced, rapid changes in motion while benefiting from LSTM's ability to interpret sequences over time.

According to the comprehensive experimental results, MobileNetV2 + LSTM showed commendable performance in dynamically active phases, particularly the Impact (I) phase, where it achieved high accuracy as depicted in the results figure 6.5. This phase, characterized by significant and rapid motion, aligns well with the model's strengths in capturing dynamic changes. However, challenges arose in static phases like Address (A) and Finish (F), where the model struggled significantly. These phases require high sensitivity to subtle, less pronounced movements, and this configuration did not effectively capture these nuances. The difficulty in detecting minimal movements in these static phases highlights a fundamental limitation of this model configuration in dealing with scenarios where motion is minimal or absent.

Building upon the baseline established by MobileNetV2 + LSTM, the study progressed to testing a more complex model, ResNet50, paired with LSTM. This model was chosen for its deeper network architecture, which is theoretically capable of extracting more complex

features from the data. ResNet50 is designed to perform deep learning tasks with enhanced depth and sophistication, potentially improving the accuracy of detecting various phases of golf swings. The integration with LSTM was intended to complement the deep feature extraction capabilities of ResNet50 with robust temporal analysis, thereby creating a potent combination for complex activity recognition.

The performance of ResNet50 + LSTM, as evidenced in the experimental data, marked a noticeable improvement in dynamically active phases such as Mid-Downswing (MD) and Impact (I), areas where more complex movements occur. This enhancement is illustrated in the results figure 6.5, highlighting the model's capability to handle complex motion dynamics more effectively than the simpler MobileNetV2 + LSTM configuration. However, this model, much like its predecessor, did not exhibit significant advancements in addressing the static phases. Despite the increased depth and complexity of the network, there was no substantial enhancement in phases characterized by minimal motion. This underscores a critical insight: deeper network architectures alone may not suffice in addressing the unique challenges posed by static environments in sports analytics.

The comparative analysis between ResNet50 + LSTM and MobileNetV2 + LSTM reveals that while the former enhances capabilities in handling dynamic phases due to its deeper learning features, both models exhibit similar deficiencies in static phases. This observation highlights the necessity for specialized mechanisms beyond mere depth in network architecture to effectively address the challenges presented by less dynamic or static environments.

The evolution of model configurations continued with the integration of MobileNetV3 with LSTM. MobileNetV3 represents an advancement over its predecessor by incorporating architectural optimizations that enhance both efficiency and performance. These optimizations include the use of more advanced activation functions and layers specifically designed to maximize processing efficiency and model accuracy. The expectation was that these enhancements would allow MobileNetV3 to leverage better feature extraction capabilities while maintaining computational efficiency, making it an ideal candidate for complex sports analytics tasks.

However, the results indicated that while MobileNetV3 + LSTM performed comparably to ResNet50 + LSTM in most dynamic phases, it did not show significant improvements in static phases, as shown in figure 6.5. This finding suggests that while architectural advancements in MobileNetV3 aid in maintaining a balance between performance and computational demands, they do not specifically target the nuanced challenges of recognizing static or subtle motions. This realization points to a performance plateau reached by conventional architectures, underscoring the need for integrating additional mechanisms, such as attention modules, to breach this barrier and enhance performance in static phase detection.

The final model configuration tested in this study was MobileNetV3 + CBAM + LSTM, which incorporates the Convolutional Block Attention Module (CBAM) into the existing framework. CBAM is an advanced attention mechanism designed to enhance the model's focus on the most relevant features within each frame by applying spatial and channel-wise attention. This focus is particularly crucial for analyzing subtle features in static phases, where traditional models falter. The hypothesis was that by directing the model's attention more precisely, CBAM would enable more accurate detection of static phases, overcoming previous limitations.

Experimental data supported this hypothesis, showing a significant improvement across all phases with the MobileNetV3 + CBAM + LSTM model, particularly in static phases such as Address (A) and Finish (F), where earlier models underperformed. The results, detailed

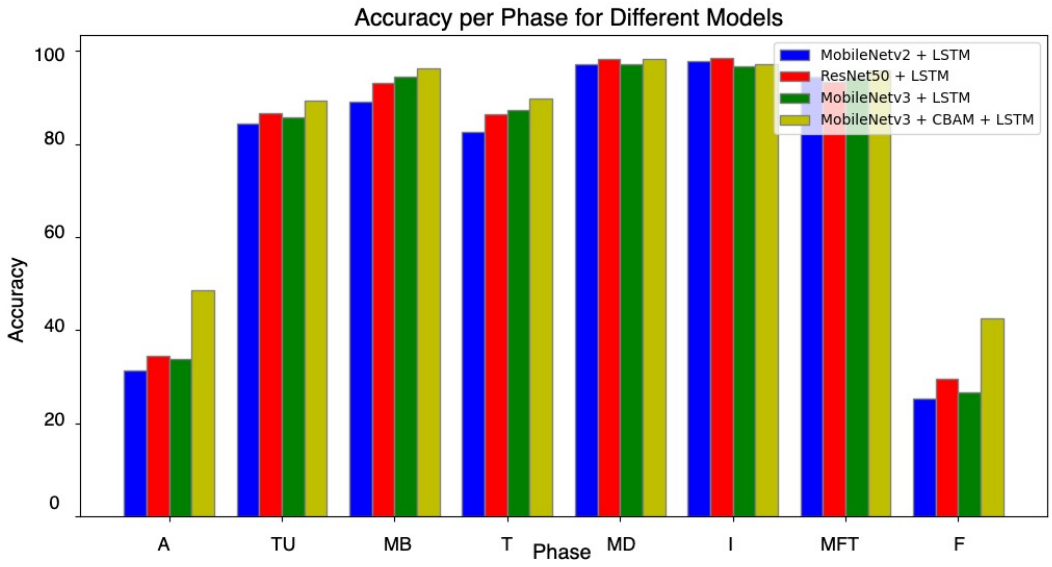

Figure 6.5: Comparative accuracy of different models across various phases of the golf swing.

in figure 6.5, illustrate the marked enhancements brought about by integrating CBAM, providing the necessary focus on subtle cues that lead to higher accuracy. This improvement demonstrates CBAM's efficacy in specifically addressing the deficiencies noted in static phase detection, thereby providing a comprehensive solution that markedly advances the field of sports analytics.

6.5.1 Research Question and Hypotheses Revisited

The central research question of this thesis was: How do advanced neural network architectures, especially those incorporating attention mechanisms, enhance golf swing analysis? This overarching question was dissected through the lens of three distinct hypotheses:

- **Hypothesis H1:** Advanced neural network architectures improve the feature extraction capabilities from video data, leading to more accurate detection of golf swing phases.
- **Hypothesis H2:** The incorporation of attention mechanisms significantly increases the precision of phase detection by focusing analysis on the most relevant features and minimizing the influence of background noise.
- **Hypothesis H3:** Integrating these advanced technologies enhances the reliability and consistency of golf swing analysis across different environments and swing types.

Hypothesis H1: Improved Feature Extraction

Our initial experiments with MobileNetV2 coupled with LSTM served as the baseline, leveraging the model's renowned computational efficiency. This configuration, designed for rapid processing on mobile devices, exhibited commendable performance, particularly in dynamically active phases such as the Impact phase, where rapid motion is predominant. The empirical results demonstrated high accuracy in these phases, supporting Hypothesis H1 by confirming the model's enhanced feature extraction capabilities.

Hypothesis H2: Incorporation of Attention Mechanisms

The integration of CBAM with MobileNetV3 marked a significant improvement in model performance, particularly in static phases such as Address and Finish. These phases, characterized by minimal motion, previously presented substantial detection challenges, which were effectively mitigated by the focused attention on relevant features. The accuracy improvements in these phases are illustrated in the comparative accuracy Figure. 6.5, substantiate Hypothesis H2.

Hypothesis H3: Reliability and Consistency

The consistency of performance enhancements across various experimental setups and swing types underscores the effectiveness of integrating advanced architectures and attention mechanisms. This consistent improvement across different environments supports Hypothesis H3, indicating an enhancement in the reliability and consistency of the golf swing analysis facilitated by these technological integrations.

The comparative analysis across models revealed differentiated capabilities in handling dynamic versus static phases. While the MobileNetV2 + LSTM and ResNet50 + LSTM configurations showed proficiency in dynamic phases, their performance in static phases was lacking. This gap was effectively bridged by the MobileNetV3 + CBAM + LSTM configuration, which excelled in static phase detection due to the targeted attention mechanisms of CBAM. This suggests that while deeper network architectures provide substantial benefits in handling complex motions, the precise detection of subtle movements in static phases requires the integration of attention-based models.

The findings from this study not only enhance the understanding of model performances in sports analytics but also set a foundation for future research into more sophisticated hybrid models combining convolutional and recurrent architectures with attention mechanisms for real-time analysis. The successful application of CBAM in this context opens avenues for exploring other attention mechanisms that could offer similar or greater enhancements in model accuracy and efficiency.

The exploration of these advanced neural network configurations in golf swing analysis has profound implications for coaching and athlete training, offering tools that provide nuanced insights into athlete performance that were previously unattainable. Further studies are encouraged to explore the scalability of these models in other sports analytics applications and to refine the attention mechanisms for even more granular feature discernment.

6.6 Comparative Performance Analysis of Our Approach Against SwingNet

In this section, we conduct a comprehensive comparison between our proposed model and SwingNet, the latter serving as a benchmark in golf swing phase detection. This analysis is designed to underline the strengths and improvements our model brings to SwingNet across a spectrum of performance metrics critical for assessing the efficacy of video-based sports analytics.

The evaluation metrics, Address (A), Top-Up (TU), Mid-Backswing (MB), Top (T), Mid-Downswing (MD), Impact (I), Mid-Follow-Through (MFT), Finish (F), and Overall Phase Classification Error (PCE)—have been selected to encompass the full range of dynamics in a golf swing. These metrics provide insights into each model's capability to identify and analyze different golf swing phases accurately, thus evaluating their overall effectiveness and

error rates.

The comparison includes:

- **SwingNet-160 (slow-motion and real-time configurations):** As the established benchmark, SwingNet-160 is tested in both slow-motion and real-time scenarios to assess its adaptability to varying video playback speeds, providing a foundation for performance benchmarks.
- **Our Approach (MobileNetV3 + CBAM + LSTM):** This model represents the culmination of integrating advanced neural network architectures with attention mechanisms, aimed at significantly enhancing both the accuracy and efficiency of swing phase detection by concentrating on the most critical features in video frames.

Additionally, to illustrate the broad spectrum of current capabilities in this domain, we include results from other configurations like ResNet50 + LSTM and MobileNetv3 + LSTM. These models are well-regarded for their spatial and temporal feature extraction capabilities from video data, providing a comprehensive context for evaluating our approach.

The performance of each model across the selected metrics is summarized in the following table (6.1). This side-by-side comparison facilitates a critical evaluation of our model's performance against the established benchmark, highlighting areas for improvement and the potential for real-world application in coaching and sports technology.

Table 6.1: Comparison of Different Models

Model	**A**	**TU**	**MB**	**T**	**MD**	**I**	**MFT**	**F**	**PCE**
SwingNet-160 (slow-motion)	23.5	80.7	84.7	75.7	97.8	98.3	98.0	21.5	72.5
SwingNet-160 (real-time)	38.7	87.2	92.1	90.8	98.3	98.4	97.2	30.7	79.2
SwingNet-160	31.7	84.2	88.7	83.9	98.1	98.4	97.6	26.5	76.1
ResNet50 + LSTM	34.5	86.7	93.1	86.5	98.2	98.5	93.4	29.5	77.55
MobileNetv3 + LSTM	33.7	85.7	94.5	87.2	97.1	96.7	94.3	26.6	76.97
Our Approach	48.6	89.3	96.2	89.8	98.3	97.2	95.8	42.6	82.22

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis explored four advanced deep-learning configurations: MobileNetV2 + LSTM (SwingNet), ResNet50 + LSTM, MobileNetV3 + LSTM, and MobileNetV3 + CBAM + LSTM. Each model was evaluated for its efficacy in detecting various phases of golf swings, with particular attention to their performance in dynamic and static phases. This investigation revealed distinct strengths and limitations of each configuration, providing insightful reflections on the nuanced capabilities of modern neural network architectures within the context of sports analytics.

SwingNet provided a robust baseline, showing high efficiency in processing dynamic activities, but struggled with static phase detection, where subtle and minimal movements dominate. The adoption of ResNet50 + LSTM improved dynamic phase detection due to its more profound and complex architecture; however, it also demonstrated that simply increasing network depth is insufficient for enhancing detection in static phases.

Further experiments with MobileNetV3 + LSTM improved computational efficiency and performance, though they did not significantly surpass previous models in static phase detection. Introducing the Convolutional Block Attention Module (CBAM) with MobileNetV3 and LSTM was a pivotal enhancement, showing marked improvements across all phases, particularly in the less dynamic swing phases. This integration underscores the transformative potential of attention mechanisms in increasing the sensitivity and accuracy of neural networks, especially in scenarios traditionally challenging for standard models.

The results of this study emphasize the importance of tailored model configurations and specialized enhancements to surpass the inherent limitations of conventional deep learning architectures. They highlight the critical necessity of aligning model capabilities with the intricate demands of complex activity recognition tasks, such as golf swing analysis, to adeptly capture the complete spectrum of dynamic and static movements.

In collaboration with Initial Force AS, this research advances the theoretical aspects of neural networks in sports analytics. It has the potential to offer practical applications that can significantly enhance real-time analysis and feedback systems in sports performance technology. The improvements in golf swing analysis can potentially improve coaching techniques and athlete performance, aligning with Initial Force AS's mission to make sophisticated analytics accessible to a broader audience.

As this study concludes, reflecting on its implications for Initial Force AS, with whom this research was conducted, is crucial. The integration of CBAM with MobileNetV3 + LSTM could be applied to Initial Force AS's data. This enhancement could elevate the company's

software applications, offering more accurate, real-time feedback to athletes and coaches and reinforcing the company's position as a leader in innovative sports technology. This collaboration showcases how academic research has the potential to translate into substantial commercial and practical benefits, underscoring the value of such partnerships in driving the field of sports analytics.

7.2 Future Work

The promising outcomes achieved with the CBAM integration open several potential avenues for future research that could further augment the utility of deep learning in sports analytics, extending its applicability beyond the current scope:

- **Exploration of Additional Attention Mechanisms:** Building on the success of CBAM, future research could investigate other attention mechanisms, such as self-attention and transformer models, adapted for video analysis to further boost performance in dynamic and static conditions.
- **Hybrid Architectures:** Investigating hybrid models that amalgamate the strengths of different neural network architectures, such as combining features from both ResNet and MobileNet with advanced attention modules, might yield more robust models capable of detecting complex motions across various sports.
- **Real-Time Analysis Implementation:** Exploring the deployment of these advanced models in real-time analysis scenarios offers an opportunity to significantly impact sports training and performance feedback systems, emphasizing the optimization of model efficiency and processing speed to enable instant feedback.
- **Enhanced Data Collection and Labeling Techniques:** Advancing data collection methods and developing more sophisticated data labeling strategies could improve model training and refinement. Collaborating with sports scientists and professional coaches would ensure the data is representative and comprehensive, aiding in developing more accurate and reliable models.

Bibliography

[1] Md Ariful Islam Anik et al. "Activity recognition of a badminton game through accelerometer and gyroscope." In: *2016 19th International Conference on Computer and Information Technology (ICCIT)*. IEEE. 2016, pp. 213–217.

[2] Gonzalo Bailador et al. "Real time gesture recognition using continuous time recurrent neural networks." In: *2nd international ICST conference on body area networks*. 2007.

[3] Benoit Bideau et al. "Using virtual reality to analyze sports performance." In: *IEEE Computer Graphics and Applications* 30.2 (2009), pp. 14–21.

[4] Peter Blank et al. "miPod 2: a new hardware platform for embedded real-time processing in sports and fitness applications." In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 2016, pp. 881–884.

[5] Qinglei Cao et al. "SGDB: a sports gene database for visualization of sports effects on human skeletal muscle gene expression." In: *IEEE Access* 8 (2020), pp. 20557–20562.

[6] 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*. 2017, pp. 6299–6308.

[7] Stephen-Mark Cooper et al. "A simple statistical method for assessing the reliability of data entered into sport performance analysis systems." In: *International Journal of Performance Analysis in Sport* 7.1 (2007), pp. 87–109.

[8] Joseph Costello et al. "Use of thermal imaging in sports medicine research: a short report: short article." In: *International SportMed Journal* 14.2 (2013), pp. 94–98.

[9] Gonçalo Dias and Micael S Couceiro. *The science of golf putting: A complete guide for researchers, players and coaches*. Springer, 2015.

[10] Jeffrey Donahue et al. "Long-term recurrent convolutional networks for visual recognition and description." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 2625–2634.

[11] Mingtao Dong et al. "HAR-Net: Fusing deep representation and hand-crafted features for human activity recognition." In: *Signal and Information Processing, Networking and Computers: Proceedings of the 5th International Conference on Signal and Information Processing, Networking and Computers (ICSINC)*. Springer. 2019, pp. 32–40.

[12] E Fenil et al. "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM." In: *Computer Networks* 151 (2019), pp. 191–200.

[13] Nicolas Gehrig, Vincent Lepetit, and Pascal Fua. "Visual Golf Club Tracking for Enhanced Swing Analysis Tools." In: *Proceedings of the British Machine Conference*. 2003, pp. 47–1.

[14] Indrajeet Ghosh, Sreenivasan Ramasamy Ramamurthy, and Nirmalya Roy. "Stancescorer: A data driven approach to score badminton player." In: *2020 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops)*. IEEE. 2020, pp. 1–6.

[15] Silvio Giancola et al. "Soccernet: A scalable dataset for action spotting in soccer videos." In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2018, pp. 1711–1721.

[16] Heiner Gillmeister. "Golf on the Rhine: On the origins of golf, with sidelights on polo." In: *The International Journal of the History of Sport* 19.1 (2002), pp. 2–30.

[17] Rohit Girdhar et al. "A better baseline for ava." In: *arXiv preprint arXiv:1807.10066* (2018).

[18] Sojeong Ha and Seungjin Choi. "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors." In: *2016 international joint conference on neural networks (IJCNN)*. IEEE. 2016, pp. 381–388.

[19] Kaiming He et al. "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[20] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory." In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[21] HM Sajjad Hossain, Md Abdullah Al Hafiz Khan, and Nirmalya Roy. "SoccerMate: A personal soccer attribute profiler using wearables." In: *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE. 2017, pp. 164–169.

[22] Andrew Howard et al. "Searching for mobilenetv3." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1314–1324.

[23] Wenbo Huang et al. "Shallow convolutional neural networks for human activity recognition using wearable sensors." In: *IEEE Transactions on Instrumentation and Measurement* 70 (2021), pp. 1–11.

[24] Dana Hughes and Nikolaus Correll. "Distributed convolutional neural networks for human activity recognition in wearable robotics." In: *Distributed Autonomous Robotic Systems: The 13th International Symposium*. Springer. 2018, pp. 619–631.

[25] Patria A Hume, Justin Keogh, and Duncan Reid. "The role of biomechanics in maximising distance and accuracy of golf shots." In: *Sports medicine* 35 (2005), pp. 429–449.

[26] Andrej Karpathy et al. "Large-scale video classification with convolutional neural networks." In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014, pp. 1725–1732.

[27] Myeongsub Kim and Sukyung Park. "Golf swing segmentation from a single IMU using machine learning." In: *Sensors* 20.16 (2020), p. 4466.

[28] Ayca Kirimtat et al. "FLIR vs SEEK thermal cameras in biomedicine: comparative diagnosis through infrared thermography." In: *BMC bioinformatics* 21 (2020), pp. 1–10.

[29] Kyeong-Ri Ko and Sung Bum Pan. "CNN and bi-LSTM based 3D golf swing analysis by frontal swing sequence images." In: *Multimedia Tools and Applications* 80.6 (2021), pp. 8957–8972.

[30] Young-Hoo Kwon et al. "Validity of the X-factor computation methods and relationship between the X-factor parameters and clubhead velocity in skilled golfers." In: *Sports Biomechanics* 12.3 (2013), pp. 231–246.

[31] Ben L Langdown, Matt Bridge, and Francois-Xavier Li. "Movement variability in the golf swing." In: *Sports Biomechanics* 11.2 (2012), pp. 273–287.

[32] Song-Mi Lee, Sang Min Yoon, and Heeryon Cho. "Human activity recognition from accelerometer data using Convolutional Neural Network." In: *2017 ieee international conference on big data and smart computing (bigcomp)*. IEEE. 2017, pp. 131–134.

[33] Junhong Li et al. "RETRACTED: Neutrosophy theory based visualization report of sports news data." In: *International Journal of Electrical Engineering & Education* 60.1_suppl (2023), pp. 1946–1956.

[34] Chen-Chieh Liao, Dong-Hyun Hwang, and Hideki Koike. "Ai golf: Golf swing analysis tool for self-training." In: *IEEE Access* 10 (2022), pp. 106286–106295.

[35] Mingqi Lv, Wei Xu, and Tieming Chen. "A hybrid deep convolutional and recurrent neural network for complex activity recognition using multimodal sensors." In: *Neurocomputing* 362 (2019), pp. 33–40.

[36] William McNally et al. "Golfdb: A video database for golf swing sequencing." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2019, pp. 0–0.

[37] Abdulmajid Murad and Jae-Young Pyun. "Deep recurrent neural networks for human activity recognition." In: *Sensors* 17.11 (2017), p. 2556.

[38] Keiron O'shea and Ryan Nash. "An introduction to convolutional neural networks." In: *arXiv preprint arXiv:1511.08458* (2015).

[39] Francisco Javier Ordóñez and Daniel Roggen. "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition." In: *Sensors* 16.1 (2016), p. 115.

[40] Joe Pavitt, Dave Braines, and Richard Tomsett. "Cognitive analysis in sports: Supporting match analysis and scouting through artificial intelligence." In: *Applied AI Letters* 2.1 (2021), e21.

[41] Basilio Pueo. "High speed cameras for motion analysis in sports science." In: *Journal of Human Sport and Exercise* 11.1 (2016), pp. 53–73.

[42] Matevž Pustišek et al. "The role of technology for accelerated motor learning in sport." In: *Personal and Ubiquitous Computing* 25 (2021), pp. 969–978.

[43] Muhammad Rafiq et al. "Scene classification for sports video summarization using transfer learning." In: *Sensors* 20.6 (2020), p. 1702.

[44] Daniele Ravi et al. "Deep learning for human activity recognition: A resource efficient implementation on low-power devices." In: *2016 IEEE 13th international conference on wearable and implantable body sensor networks (BSN)*. IEEE. 2016, pp. 71–76.

[45] Charissa Ann Ronao and Sung-Bae Cho. "Human activity recognition with smartphone sensors using deep learning neural networks." In: *Expert systems with applications* 59 (2016), pp. 235–244.

[46] Mark Sandler et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[47] Walter Thomas Schmid. *Golf as Meaningful Play: a philosophical guide*. Lexington Books, 2017.

[48] Karen Simonyan and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." In: *Advances in neural information processing systems* 27 (2014).

[49] Tim Steels et al. "Badminton activity recognition using accelerometer data." In: *Sensors* 20.17 (2020), p. 4685.

[50] Bastien Vanderplaetse and Stephane Dupont. "Improved soccer action spotting using both audio and video streams." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 896–897.

[51] Eric C Wilson. "The effectiveness of adult education principles in teaching the golf swing." PhD thesis. Capella University, 2007.

[52] Sanghyun Woo et al. "Cbam: Convolutional block attention module." In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[53] Erwin Wu et al. "VR alpine ski training augmentation using visual cues of leading skier." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 878–879.

[54] Jianbo Yang et al. "Deep convolutional neural networks on multichannel time series for human activity recognition." In: *Ijcai*. Vol. 15. Buenos Aires, Argentina. 2015, pp. 3995–4001.

[55] Shuainan Ye et al. “Shuttlespace: Exploring and analyzing movement trajectory in immersive visualization.” In: *IEEE transactions on visualization and computer graphics* 27.2 (2020), pp. 860–869.

[56] Serena Yeung et al. “End-to-end learning of action detection from frame glimpses in videos.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2678–2687.

[57] Ming Zeng et al. “Convolutional neural networks for human activity recognition using mobile sensors.” In: *6th international conference on mobile computing, applications and services*. IEEE. 2014, pp. 197–205.

[58] Longfei Zheng et al. “Application of IndRNN for human activity recognition: The Sussex-Huawei locomotion-transportation challenge.” In: *Adjunct proceedings of the 2019 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2019 ACM international symposium on wearable computers*. 2019, pp. 869–872.