Genreating Phases of Play In Rugby Union Using Recurrent Neural Networks

Presented by:

Thomas Edley
EDLTHO001

Graham Davies
DVSGRA012

Supervisor:

Dr Neil Watson

Submitted to the Department of Statistical Sciences at the University of Cape Town in partial fulfilment of the academic requirements for a Bachelor of Science Honours in Statistical Sciences

August 9, 2021

Keywords: Recurrent Neural Networks, Long Short-Term Memory, Sequence Generation, Phases

Declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
2. I have used the IEEE convention for citation and referencing. Each contribution to, and quotation in, this report from the work(s) of other people has been attributed, and has been cited and referenced.
3. This report is my own work.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as their own work or part thereof.

Signature: ....
T. Edley

Signature: ....
G. Davies

Date: August 9, 2021

Acknowledgements

Abstract

Statistical analysis has become an important aspect in the domain of sports. Whilst most sporting analysis is based around aggregate performance measures over a fixed time interval, the sequential nature of sports is often overlooked[1]. There have been studies that have used the sequential nature of sport to predict the outcome of games in various sports, as well as other aspects of the game. Due to the sequential nature of Rugby Union, there is an opportunity to investigate whether statistical models can be used to generate sequences of play that are similar to that of a normal game of Rugby. The primary aim of this project is to determine whether one can use recurrent neural networks (RNN), in particular, long short-term memory (LSTM) networks, to simulate a game of rugby union. This will be achieved by training an RNN that is able to generate realistic phases of play using the sequences of actions by players as input data. Due to the spatio-temporal nature of the data, we aim to use the field locations of these phases, as well as the team and players involved as additional input data. If time allows, after achieving the primary objective we will also consider secondary objectives, such as investigating the sequences generated by winning and losing teams to determine if there are any significant differences in the output achieved.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The sport of Rugby Union is based around two teams of 15 players each competing for the possession of the ball and trying to move closer to the opposition's goal-line. Points are achieved by placing the ball in these goal-line areas, known as scoring a try, or by taking penalty kicks and drop goals, where a player will attempt to kick the ball through the posts situated in the middle of the oppositions goal-line. Teams with possession move the ball through phases. These phases are sequences of play that involve actions performed by the players on each team in an attempt to move closer towards the oppositions goal-line. Each phase can consist of multiple actions. Some examples of these actions are a 'carry', 'tackle' or 'pass'. Phases typically occur between 'breakdowns'. A 'breakdown' typically occurs when a player is tackled to the ground by a player on the opposition team. After being brought to the ground, multiple players from each team compete for the ball in a situation that is known as a 'ruck'. The movement of the ball from the ruck indicates the start of a new phase. Breakdowns also occur in other situations, such as the ball being kicked out or a penalty being awarded.

1.2 Objective

The primary objective of this project is to determine whether one can use RNNs, in particular, LSTMs, to simulate a game of rugby union.

1.3 Outline

This section will outline the remainder of the project:

Chapter 2 is a review of relevant literature that focuses on the use of RNNs in different domains, more importantly in the domain for sport sequence generation. In Chapter 3, section 3.1 and Section 3.2 describe Recurrent Neural Networks and Long Short-Term Memory Networks respectively, the models used for this project. Chapter 4 describes the data that is used in the project, including how it has been collected and how it has been prepared. Chapter 5 gives an overview of how the data has been modelled using LSTM networks to achieve the desired outcome. Chapter 6 is where the results of the project will be outlined, with these results being discussed in more detail in Chapter 7 and Chapter 8.

Chapter 2

Literature Review

This section focuses on a variety of literature relating to our topic. We noted that the use of RNNs for sequence generation in rugby union has limited coverage. Following this, we split up the topic and focused on reading literature on RNNs/LSTMs in sequence generation as well as reading up on the use of RNNs in sports-related studies, more specifically, invasion sports.

2.1 Sequence Generation

LSTMs have most commonly been used in the domains of natural language processing and image recognition. Graves (2014) [2] wrote the seminal paper on RNNs, in which he uses an LSTM model to develop predictions on handwriting. Positional ”XY” data of the handwriting taken from a smart whiteboard was used to train an LSTM network without any preprocessing of the collected data. This model was not only able to predict the next letter or word, but could generate sequences based on different initial starting points. [2] Our study draws similarities to this method, but applied to the field of sports tracking. Another area where sequence generation has been used is that of human trajectory prediction, which refers to the prediction of human movements based on past positional data. Alahi et al. (2016) train an LSTM model that is able to learn general human movement and predict their future trajectories. They found that the LSTM model outperforms state-of-the-art prediction methods. [3]

In Karpathy's blog [4], the effectiveness of RNNs is demonstrated. He states that the glaring limitation with Convolutional Neural Networks is that their API is too constrained. They accept a fixed-sized vector as an input, and produce a fixed-sized vector as an output. [4] Additionally, Convolutional Neural Networks are limited by a fixed amount of computational steps. [4] It is noted that sequence regime of operation is much more powerful compared to fixed networks. [4] The fundamentals of RNNs is demonstrated and is followed by five examples of the capabilities of RNNs. Included in these examples is the of RNNs to generate Paul Graham essays, Shakespeare, Wikipedia pages, Latex (Algebraic Geometry) and generating baby names.

Garcia-Valencia et al. [5] studied sequence generation using deep recurrent networks in music. They evaluated three different types of memory cells mechanisms and analysed their performance music sequence generation. They performed experiments where they compared Neural Architecture Search (NAS), Update Gate RNN (UGRNN) and LSTMs cells. The LSTM cells were found to be the best as the consecutive repetition of notes is almost null, which demonstrated excellent learning capacity to keep the scale. [5]

Neil et al. [6] demonstrates Phased LSTMs. They state that current RNN models are poorly suited to process irregularly sampled data triggered by events generated in continuous time by sensors or other neurons. [6] They introduce the Phased LSTM model. This extended on the LSTM model by adding a new time gate. They found that the model greatly improves the performance of LSTMs in standard LSTM applications. [6]

The opening and closing of this gate is controlled by an independent rhythmic oscillation specified by three parameters. The real-time period of the oscillation, the ratio of the duration to the "open" phase and phase shift of the oscillation to each Phased LSTM cell. This creates a rhythmic periodicity. [6] The rhythmic nature of Phased LSTMs allows the model to converge more quickly than regular LSTMs. [6] It also allows a shortcut to the past for gradient backpropagation, which means accelerated training. The results achieved in this paper illustrate the ability to discriminate rhythmic signals to learn long-term memory traits. [6]

2.2 Sport Related Studies

There has been little coverage of statistical analysis in rugby union. We subsequently read literature on the use of RNNs and LSTMs in other sports.

Zhang et al. (2020) used a bidirectional LSTM for the e-sport Defense of the Ages 2 (DOTA2). [7] The purpose was to use LSTMs to improve their lineup recommendations. They used a continuous bag of words model (CBOW) to predict the context of a word in a sentence. With this, a word is transformed into a hero, a sentence a lineup and a word vector into a hero vector. Using this improved Bi-LSTM model, they were able to achieve an average accuracy rate of five recommended heroes of 67.74%. [7]

Vaswani et al. (2020) proposed an autoencoder based machine learning pipeline approach to simulate the champions league from 2014 to 2020. [8] They noted that there had previously been little focus on the intricacies of football that might be of interest. They created three handcrafted features, a home/away index, a form index experience. Using these features, they did not only predict the results and simulate the Champions League, but they also predicted intricate statistics like corners, goal scorers, possession and passes. This suggests that in future work can be done by weighting the time of the matches so that the older matches will have a lower weighting than newer matches. [8]

Goddijn et al. (2018) compared three techniques to predict the outcomes of football matches. [9] They used simple logistic regression, a 3-layer neural network, and an LSTM. The neural network was found to be the best with a development accuracy of 51% despite the fact LSTMs could take in historical data. The LSTM model was found to have a development accuracy of 47.34% after adjustments for overfitting. [9] They suggest that future work can be done by looking at different architecture types and LSTM variations. [9]

Romijnders and Shah (2016) [10] use RNNs in the form of sequence generation to predict whether a three-point shot in a professional basketball game will be successful. They trained models that were able to learn the trajectory of a basketball without any knowledge of the physics behind it. The results of the model were compared to those of a baseline static machine learning model with a full set of features, such as angle and velocity, in addition to positional "XY" data. They found that models based simply on sequential positional data outperform a static feature-rich machine learning model in predicting whether a 3 pointer will be successful or not. The model used was a two-layered LSTM using peephole connections. [10]

A study by Verpalen (2019) [11] uses LSTMs to predict player movement in Soccer using XY data. The study focused on the prediction of player movement via object tracking from visual data such as images and videos. The results show that LSTMs are well capable of predicting the next change in the players' position. However, on longer input sequences, the predictive performance decreases. LSTMs were found to be a suitable deep learning technique for predicting movement in a game of soccer. One of the key outcomes of this study is that they were able to not only predict the next movement of individual players one at a time, but were able to predict the next positions of all the players on the field at the same time. [11]

A majority of the statistical studies done on rugby union have focused on outcomes based on key performance indicators (KPIs). Watson et al. (2020)[1], who use the same data as this project, used convolutional and recurrent neural networks to predict certain outcomes of sequences of play, based on the order of the phases and the position in the field where the action occurs. The performance of the model to that of a baseline Random Forrest. They used these results to investigate how their model could provide tactical decision-making support to rugby coaches. The result of this study was that when the sequence of actions and position of these actions on the field are used as input for a convolutional and recurrent neural network, prediction is more accurate than the baseline model. They demonstrated that the sequential nature of rugby union is an important factor in the assessment of team performance. [1]

As far as they were aware, their study was the first to consider the sequential nature of rugby union in predicting intra-game outcomes. This project draws many similarities from their study, however, we will be using the sequential nature of the data to try and simulate a game of rugby union. From the research we have done, we are not aware of any studies that have attempted to investigate the same outcome as this study does, which therefore makes this project and exciting an unique experiment.

Chapter 3

Methods

3.1 Recurrent Neural Networks

Recurrent neural networks (RNNs) are a rich class of dynamic models that can be used to model sequential data. It is the first algorithm that can remember its input, which makes it very useful for machine learning problems that involve sequential data. From Graves' seminal paper on generating sequences using recurrent neural networks [2], RNNs are similar to standard feed-forward neural networks, except that their output distribution is additionally influenced by internal memory. Sequences are generated from a trained network by iteratively sampling from the output distribution. These samples are then fed into the network as input at the next step, such that the network is taking in new input as well as input from previous iterations. RNN's have the form of a chain of repeating modules of neural networks. In Figure 2.1 below, which shows the structure of a standard RNN, the repeating module has a very simple structure with a single tanh layer [12].

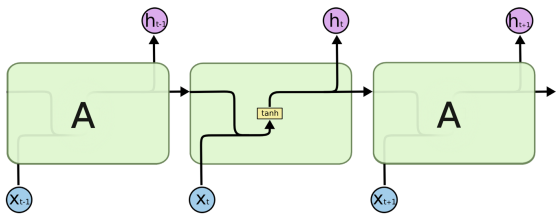

Figure 3.1: The repeating module in a standard RNN contains a single layer.

Standard RNNs are unable to information about past inputs for a long period of time [13].

3.2 Long Short-term Memory Networks

Standard RNNs are unable to information about past inputs for a long period of time (Hochreiter et al). LSTMs are extensions of RNNs that are capable of learning long-term dependencies. They were introduced by Hochreiter and Schmidhuber [13] and have been widely used since then in many domains. Unlike RNNs, they are able to remember information for long periods of time. Just like RNNs, LSTMs have a chain-like structure however the repeating module has a different structure [12]. Figure 3.2 [12] displays the structure of an LSTM network. From this, we can see that an LSTM network has four neural network layers, as opposed to RNNs which have one. These four layers are what allow LSTMs to remember information for a longer time.

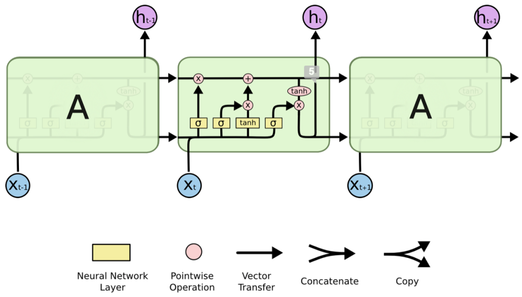

Figure 3.2: The repeating module in an LSTM contains four interacting layers.

The horizontal line that runs through the top of the LSTM module is known as the cell state, and is the key part of the LSTM. Information is added or removed from the cell state depending on the nature of the structures known as gates. Gates allow LSTMs to optionally let information through the cell state. They are composed out of a sigmoid neural net layer and a pointwise multiplication operation. They are indicated by the yellow blocks with sigma symbols on them in Figure 3.2.1. There are three different gates. The input gate determines whether new input should be let in or not, the output gate determines the impact on the output distribution at the current step and the forget gate will delete information if it is not important.

More in-depth explanation to follow.

Chapter 4

Data

The data consists of five different rugby competitions, namely, The Heineken Cup, the European Rugby Championship, Super Rugby, The Six Nations and The Rugby Championship. The years/seasons vary between competitions but range from 2013 to 2015. The data is extensive and contains 23 variables for each observation. The data for each competition is stored on a separate sheet, all of which are in the same format. The data contains positional variables such as the location on the field, time and action qualifiers. There are 29 action events, and an additional 475 action qualifiers to describe these events.

4.1 Data Collection

The data was supplied by our supervisor, but originated from Opta, a sports analytics company. The data was collected by two teams of two Opta analysts. The post-match screening was then performed by another two analysts which involved numerous accuracy checks. These analysts require training for three to six months before coding live games, and additionally, Opta monitors each analysts accuracy throughout the season by performing regular accuracy checks.

4.2 Data Preparation

At this stage of the study, we have not decided which variables we will be using. This section will be updated at a later stage.

Chapter 5

Model Fitting

Chapter 6

Results

Chapter 7

Conclusions

Chapter 8

Discussion

Bibliography

[1] N. Watson, S. Hendricks, T. Stewart, and I. Durbach, "Integrating machine learning and decsision support in tactical decision-making in rugby union," *Journal of the Operational Research Society*, 2020.

[2] A. Graves, "Generating sequences with recurrent neural networks," 2014.

[3] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: HumanTrajectory Prediction in Crowded Spaces," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–971, 2016.

[4] A. Karpathy, "The Unreasonable Effectiveness of Recurrent Neural Networks," 2015.

[5] J. G. L.-P. Sebastian Garcia-Valenciaa, Alejandro Betancourtb, "Sequence generation using deep recurrent networks and embeddings: A study case in music," 2020.

[6] D. Neil, M. Pfeiffer, and S.-C. Liu, "Phased lstm: Accelerating recurrent network training for long or event-based sequences," 2016.

[7] L. Zhang, C. Xu, Y. Gao, Y. Han, X. Du, and Z. Tian, "Improved Dota2 Lineup Recommendation Model Based on a Bidirectional LSTM," *Tsinghua Science and Technology*, vol. 25, no. 6, pp. 712–720, 2020.

[8] A. Vaswani, R. Ganguly, H. Shah, S. Ranjit, S. Pandit, and S. Bothara, "An autoencoder based approach to simulate sports games," 2020.

[9] S. Goddijn, E. Moshkovich, and R. Challa, "A Sure Bet: Predicting Outcomes of Football Matches," 2018.

[10] R. R. Rajic C. Shah, "Applying deep learning to basketball trajectories," 2016.

[11] J. Verpalen, "Predicting player movements in soccer using Deep Learning," 2019.

[12] C. Olah, "Understanding LSTM Networks," 2015.

[13] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-term Dependencies." *A Field Guide to Dynamical Recurrent Neural Networks*, 2001.

Appendix A

Link to GitHub Repository

The code used for this paper can be found in the following GitHub repository: