Match Statistics Prediction

Kürşat Öztürk
Department of Computer Engineering
METU
Ankara, Turkey
kursatozturkcs@gmail.com

Abstract

In this paper, deep learning methods are investigated to predict statistics of a football match. The dataset contains previously played matches' statistics from all over the world as long as the event has statistics in data source[1].

In this work, I have studied several type of neural networks to predict approximate values of statistics. Best accuracy achieved by convolutional network.

I.Introduction

There are several attempts on predicting the football matches' winners by using previous matches' data, but there is not exists optimum solution could predict more accurately than humans, or betting companies. Although in this industry there exists a vast amount of money, and people are putting more importance on statistics day by day, analyzing still done by area experts, manually.

This work aims to achieve a solid accuracy on both analyzing what is going to happen in match just by looking what happened previous matches played by teams.

A. Goal

This work aims to predict how teams will play in terms of statistics, by looking last matches. So determine what one can wait from a football match.

B. Constraints

Dataset consist of samples order of ten thousands, which faces the danger to easily overfitting. Choosing layer counts, parameters were way more tricky than a adequate dataset. To avoid this, samples are cutted to two halves and network have two inputs and two outputs for each half.

C. Previous Works

There are several thesis studied the prediction of the winner of a match. Petterson & Nyquist (2017)[2], used LSTM to predict result, where feeded values were from matches' 15 minute plays.

HERBINET (2018)[3] studied several machine learning methods to achieve same goal. Bosch(2018)[4] used same methods to predict NFL matches' winners, but the strongest accuracy was around %60. There does not exists a completed thesis on predicting the statistics of a match.

II.Methodology

A. Data

Dataset consist of the statistical data. Statistics contains 33 different feature that team can produce in a match. Network is feeded by 4 input, home team's and their opponents' average statistics in last 5 matches, and away team's and their opponents' average statistics in last 5 matches.

I got data from SofaScore[1], which is a live sports statistics and score streaming platform.

An example of statistical data:

{'awayPasses': 161, 'homePasses': 326, 'awayDuelWon': 29,
'homeDuelWon': 26, 'awayDuelLost': 26, 'awayOffsides': 1,
'awayRedCards': 0, 'awayThrowIns': 12, 'homeDuelLost': 29,
'homeOffsides': 0, 'homeRedCards': 0, 'homeThrowIns': 14,
'awayAerialWon': 10, 'awayFreeKicks': 3, 'awayGoalKicks': 6,
'homeAerialWon': 8, 'homeFreeKicks': 4, 'homeGoalKicks': 1,
'awayAerialLost': 8, 'homeAerialLost': 10, 'awayCornerKicks': 0,
'awayHitWoodwork': 0, 'awayShotsOnGoal': 2, 'awayYellowCards': 0,
'homeCornerKicks': 6, 'homeHitWoodwork': 0, 'homeShotsOnGoal': 2,
'homeYellowCards': 0, 'awayShotsOffGoal': 0, 'homeShotsOffGoal': 7,
'awayAttOutBoxPost': 0, 'awayKeeperSweeper': 0, 'homeAttOutBoxPost': 0,
'homeKeeperSweeper': 1, 'awayAccuratePasses': 126, 'awayBallPossession': 33,
'awayDuelWonPercent': 53, 'homeAccuratePasses': 295, 'homeBallPossession': 67,
'homeDuelWonPercent': 47, 'awayAttInBoxBlocked': 2, 'awayGoalkeeperSaves': 2,
'homeAttInBoxBlocked': 1, 'homeGoalkeeperSaves': 2, 'awayAerialWonPercent': 56,
'awayAttInsideBoxMiss': 0, 'awayAttOutBoxBlocked': 0, 'awayTotalShotsOnGoal': 4,
'homeAerialWonPercent': 44, 'homeAttInsideBoxMiss': 1, 'homeAttOutBoxBlocked': 0,
'homeTotalShotsOnGoal': 10, 'awayAttOutsideBoxMiss': 0, 'homeAttOutsideBoxMiss': 6,
'awayAttInsideBoxTarget': 1, 'homeAttInsideBoxTarget': 1, 'awayAccurateThroughBall':
0, 'awayAttOutsideBoxTarget': 1, 'awayTotalShotsInsideBox': 3,
'homeAccurateThroughBall': 0, 'homeAttOutsideBoxTarget': 1,
'homeTotalShotsInsideBox': 3, 'awayTotalShotsOutsideBox': 1,
'homeTotalShotsOutsideBox': 7, 'awayAccuratePassesPercent': 78,
'awayBlockedScoringAttempt': 2, 'homeAccuratePassesPercent': 90,
'homeBlockedScoringAttempt': 1}

Fig 2.1: Features that are used in training process.

B. Network

In this work, my approach was to separate two halves and treat them two different event that each one affects the other in a very complicated way. In this approach, I had the dilemma that should the network be fed by two separate event or should I treat as two halves altogether consists an event.

Model Type	Validation Loss
Fully Connected Network with one input/output (in one forward pass both first half and second half are fed to network)	4.3510
Convolutional Network with one input/output (in one forward pass both first half and second half are fed to network)	4.8248
Recurrent Convolutional Network with LSTM (in one forward pass only one half fed to network, each sequence consist of two halves)	4.4181
Convolutional Network with one input/output (in one forward pass only one half fed to network)	9.1915
Convolutional Network with two input/output	**4.2529**

Fig2.2: Neural Networks and their validation loss

Although in terms of validation loss fully connected layer and lstm seems pretty close to the convolutional, when I examine prediction results, convolutional network produce more reliable results than the others on reflecting the characteristic play of teams.

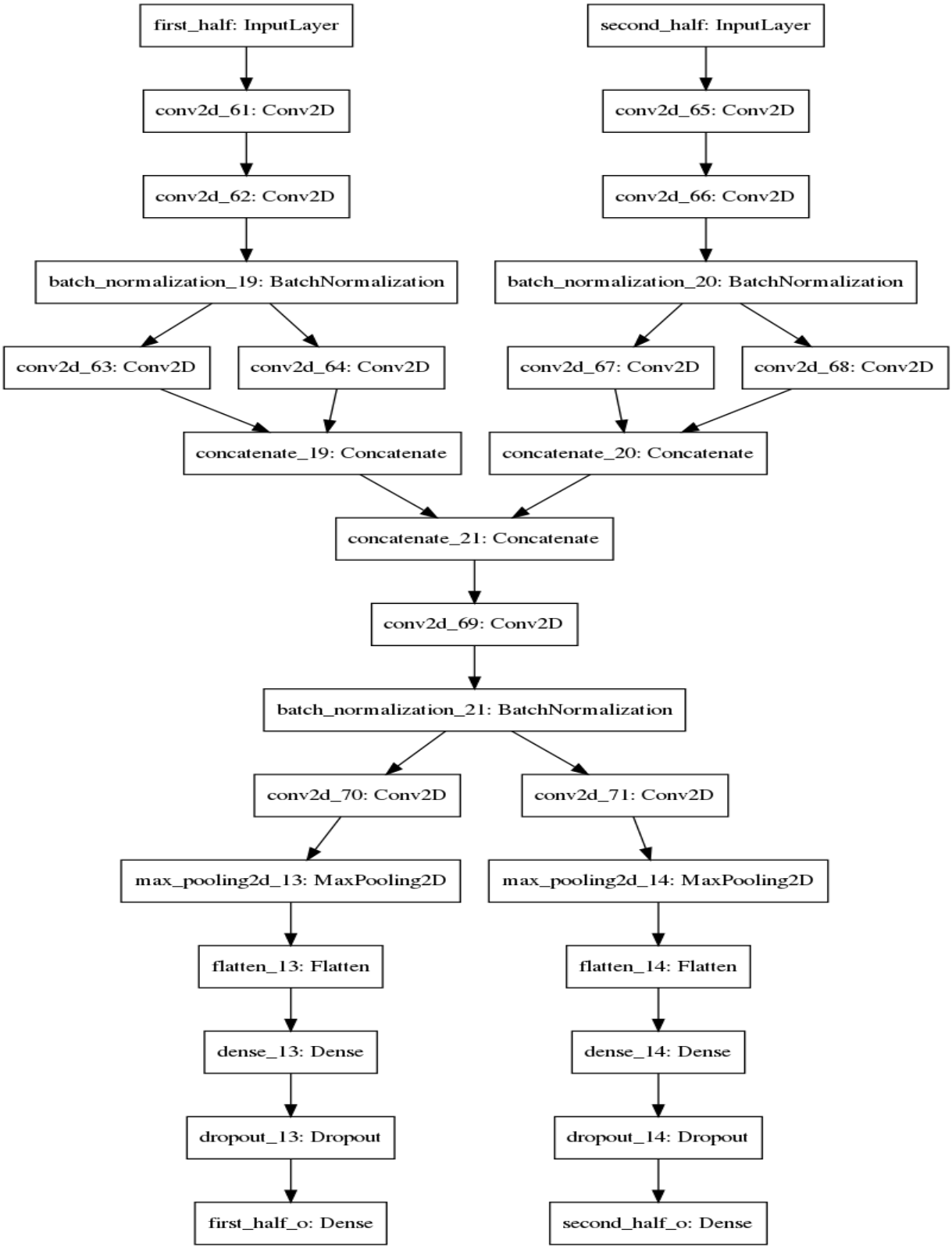

Fig 2.2: Network architecture

III. Results

Even predicting the winner is a way tricky and hard to achieve, predicting statistics has more hardness in its nature. Dataset contains ~21.000 match sample, which is to feed the network a bit tiny set. However, convolutional network achieved to learn characteristics of statistics even in a few epochs.

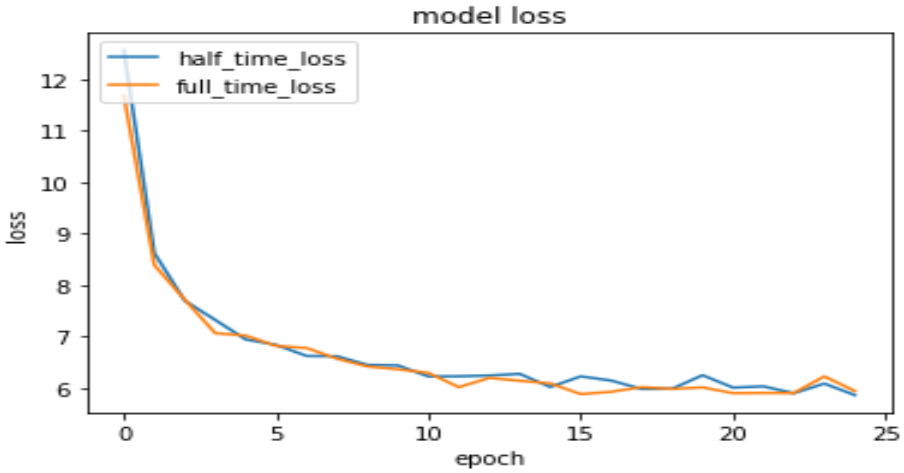

Fig 3.1: Loss curve on initializing network with a small subset of data.

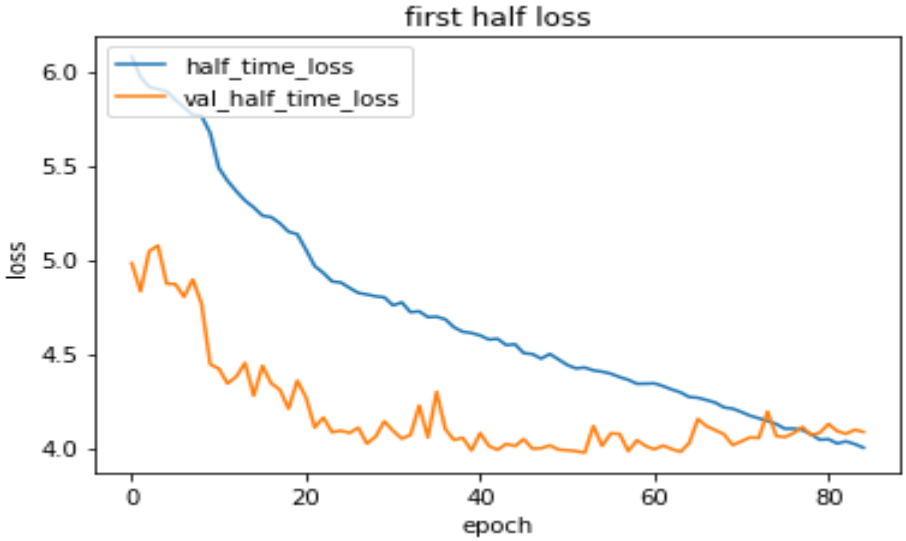

Fig 3.2: Loss curve of First half statistics output during training.

Fig 3.3: Loss curve of Second half statistics output during training.

IV. Conclusion

It cannot predict precisely; however, network learned the characteristics of statistics. With a proper dataset, it can be achieved that a Convolutional Network may achieve more precise statistics prediction just by looking last matches.

Even though predicting one person's behaviour is a tricky process, with a hard work, such events that depends on multi-people's behaviour will be predictable within a neural network in the near future.

References

1. *https://www.sofascore.com*
2. Daniel Petterson and Robert Nyquist , “Football Match Prediction using Deep Learning”, Department of Electrical Engineering Chalmers University of Technology Gothenburg, Sweden 2017
3. Corentin Herbinet, “Predicting Football Results Using Machine Learning Techniques”, Imperial College of London, 2018
4. Pablo Bosch, “Predicting the winner of NFL-games using Machine and Deep Learning.“, Vrije universiteit, Amsterdam, Nederland Research Paper Business Analytics, 2018