Minh Tran

Alex Bank

SPA DRP Writeup

Spring 2024

Deep Learning on Sports Data

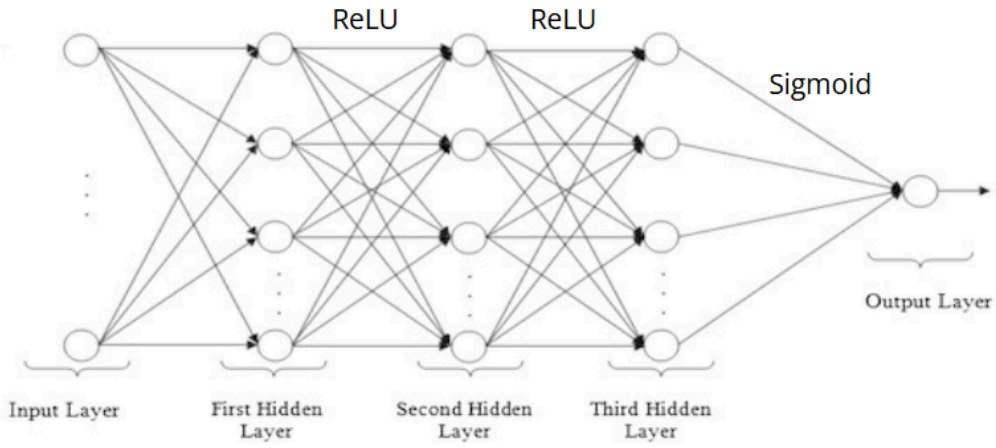
This quarter, I and another student were given the opportunity to dive into deep learning techniques and its applications to sports data analysis. More specifically, we used the data provided by the UW Women Volleyball coach, including match events data from the past 2023 season from Power Five Conferences games (ACC, Big Ten, Big 12, Pac-12, SEC). The overall goal that we set up at the beginning of the quarter was to create a useful tool that the coaches can use in the future to analyze or improve their team performance. Before this Directed Reading Program (DRP), I was not entirely familiar with the concept of neural networks, but I wanted to go deeper into this topic. Throughout the quarter, we explored architectures ranging from simple neural networks Multilayer Perceptron (MLP) to Recurrent Neural Network (RNN). We not only learned about the architecture but also the mathematical concepts behind the models we selected like activation function, gradient descent, or back propagation.

In the first few weeks of the quarter, we spent most of our time setting up, cleaning the data and I also contacted the coach asking for more data because we will need lots of data to train a neural network. I also read the paper "Estimating individual contributions to team success in women's college volleyball" to understand more about volleyball as well as others' approaches to model player contribution based on DataVolley data. To get more familiar with neural networks, I also worked through some PyTorch tutorials and the 3Blue1Brown series on neural networks. We decided our objective would be to use match information to predict the number of winning attacks by both home and away teams, as scoring an attack is a crucial part of volleyball. Each input for the model would be the event sequence of the entire game. We then started building a simple MLP model and gradually increased the complexity of the architecture to see how it performed in prediction. Although the progress looked promising at first, we

eventually hit a wall with MLPs as the test loss stopped decreasing, and we used Mean Squared Error (MSE) Loss as our loss function..

We then switched to using an RNN, hoping to better utilize the sequential nature of the data. This approach also reached a deadend as the model keeps giving us the same output regardless of the inputs when we test it on our test set. Potential reasons for this result could be that we didn't have a large enough amount of data for RNN as well as the possibility of vanishing gradient due to large sequence length and the nature of RNN. Since it seemed like we couldn't improve the model with the same input and output structure, I changed my approach. Instead of using the whole game data, for each game, use all sequences that lead to an attack (Dig → Set → Attack). My new inputs were the data from the Dig and Set timepoints, with padding where necessary. The corresponding new output would be whether the following attack would be a winning attack or not (0 or 1). By changing the approach, I now have a total of 28901 attacks (only from Pac-12) compared to only 629 games (from all Power Five Conferences) in the previous approach. I also use One-Hot Encoding to increase the number of features to 194.

The architecture of my final model is 3 hidden layers, each with 500 nodes, using ReLU activation function after the first and second layer, and using Sigmoid after the third layer. Since this is now a binary problem, the loss function that I used is Binary Cross Entropy Loss.



On the test set, I achieved a test loss of 0.3433, an accuracy of 80.64%, a precision of 85.44%, and a recall of 56.72%.

This is pretty good considering I haven't used all the available data due to computational limitations on Google Colab. In the future, I plan to refine this model with better data preprocessing and research better architectures. Moreover, I want to identify the features the model considers important to make useful suggestions to the coach and provide the coaching staff with a pre-built model they can use to analyze performance.

Overall, I had a great experience with this DRP, exploring many interesting topics in deep learning and applying them to a real-world sports analytics problem. This project not only enhanced my understanding of neural networks but also showed me their potential implications in sports. I'd also like to thank my mentor, Alex Bank, for his significant contributions to my learning in this niche area that I'm very passionate about. His guidance was invaluable, and I'm looking forward to continuing this journey and seeing where it takes me in the field of sports analytics and beyond.