Deep Learning on Sports Data

Mentees: Minh Tran, Weixuan Liu
Mentor: Alex Bank

Overview:

Part 1: Volleyball

- Basic Information of the volleyball
- Introduction to the volleyball dataset
- Some Visualizations on dataset

Part 2: Modeling

- Fundamentals on Neural Network, MLP and RNN
- Architecture of our models
- Learning Performance

Part 3: Discussion

- Impacts
- Future Work
- Takeaways

Volleyball

How to score a point

- **Attack:** An attack is recorded any time a player attempts to attack the ball into the opponent's court.
- **Winning Attack:** A successful attack that scores

Data

- The data is provided by UW Women Volleyball Coach, which include all matches in the Power 5 Conferences from the 2023 season.

  (629 games)

DVW File

Detailed information about the corresponding volleyball match

Features (86)

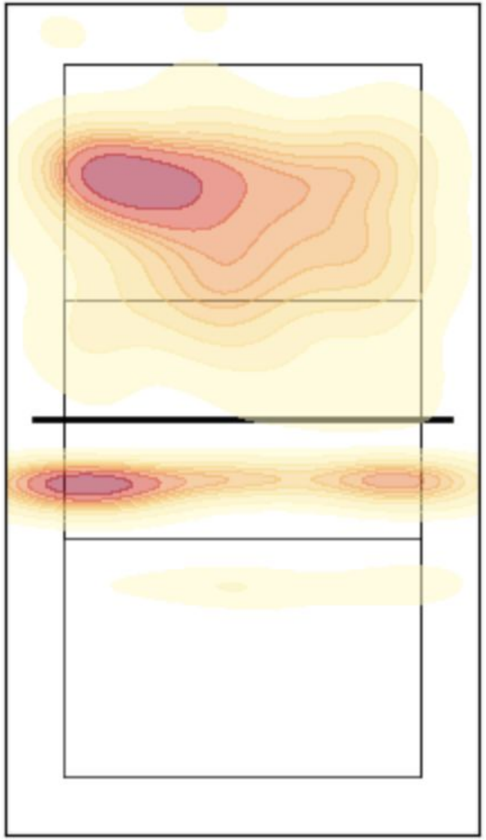
Time Steps (2100-3000)

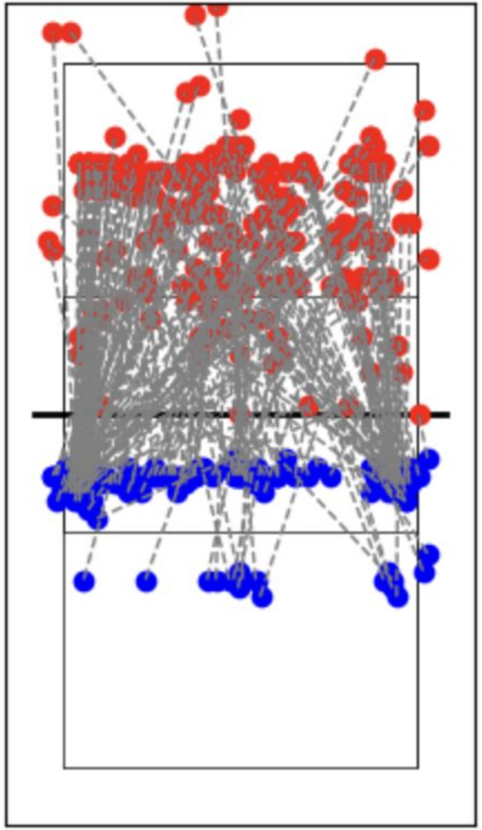
	match_id	video_file_number	video_time	code	team	player_number	player_name	player_id	skill	evaluation_code	setter_po
7	63e71130-e85b-4d86-8e9d-9d5abaa252b2	1	595	a21AT+X5¯41CH2¯00F	Oregon State University	21	Megan Sheridan	-432404	Attack	+	
12	63e71130-e85b-4d86-8e9d-9d5abaa252b2	1	603	a06AT-X6¯28AH4¯00F	Oregon State University	6	Mychael Vernon	-311736	Attack	-	
15	63e71130-e85b-4d86-8e9d-9d5abaa252b2	1	607	*10AT=X8¯96AH2¯00B	Stanford University	10	Kendall Kipp	-282546	Attack	=	
25	63e71130-e85b-4d86-8e9d-9d5abaa252b2	1	633	*17AN#CF¯26BH4¯-1F	Stanford University	17	Sami Francis	-336492	Attack	#	
36	63e71130-e85b-4d86-8e9d-9d5abaa252b2	1	663	a21AH#V6¯29DH2¯00F	Oregon State University	21	Megan Sheridan	-432404	Attack	#	
...	...	...	...	...	...	...	...	...	...	...	
1927	63e71130-e85b-4d86-8e9d-9d5abaa252b2	1	7928	a09AT#X5¯41CH2¯-9F	Oregon State University	9	Peyton Suess	-432401	Attack	#	
1939	63e71130-e85b-4d86-8e9d-	1	7962	*10AT-X5¯47AH2¯+9F	Stanford University	10	Kendall Kipp	-282546	Attack	-	

Some Visualizations

Coordinates of Volleyball End Zone for Attacks

Coordinates of Volleyball Start Zone for Attacks





Goal

- Provide helpful insights for the coach

- Leverage deep learning techniques to analyze volleyball data

Gap: The power of new emerging advanced DL

VS

Many people still use statistical methods on volleyball data

Objective of Modeling

Use some information about a match to predict the number of winning attacks of both home and away teams.

Inputs:

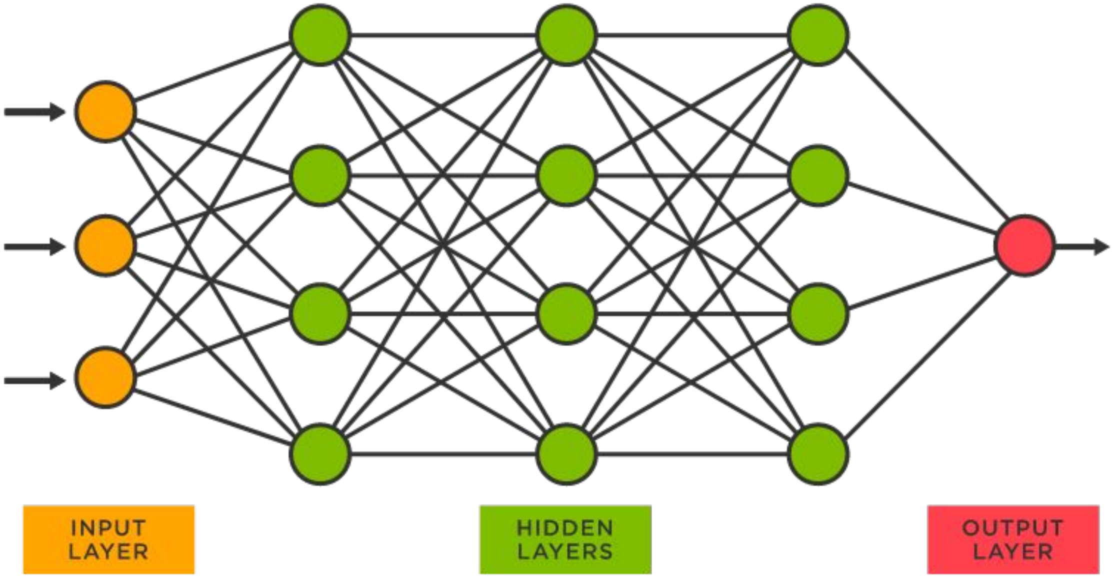
Manually select 37 features that may have potential influence on the number of winning attacks.

Output:

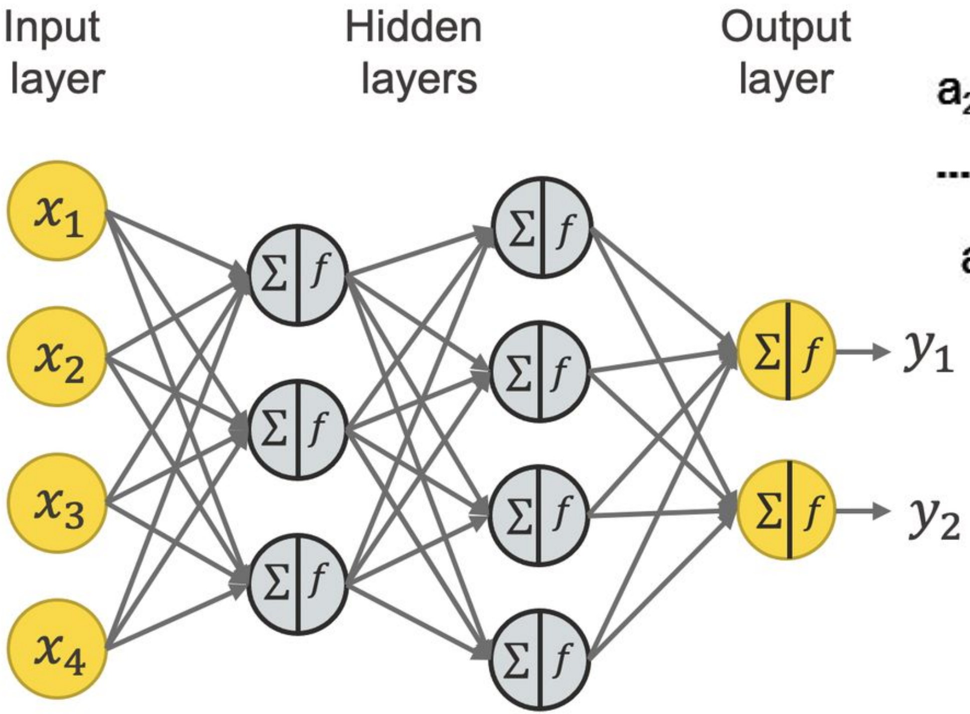
The total number of winning attacks of home team and away team.

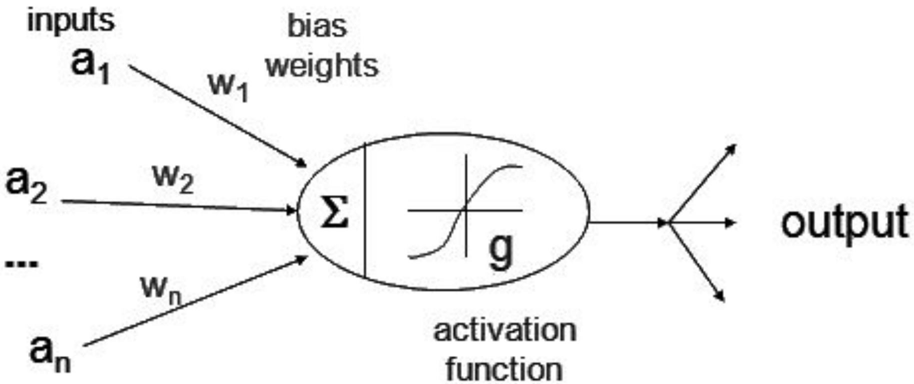
Fundamentals of Deep Learning

Neural networks are the underlying technology in deep learning

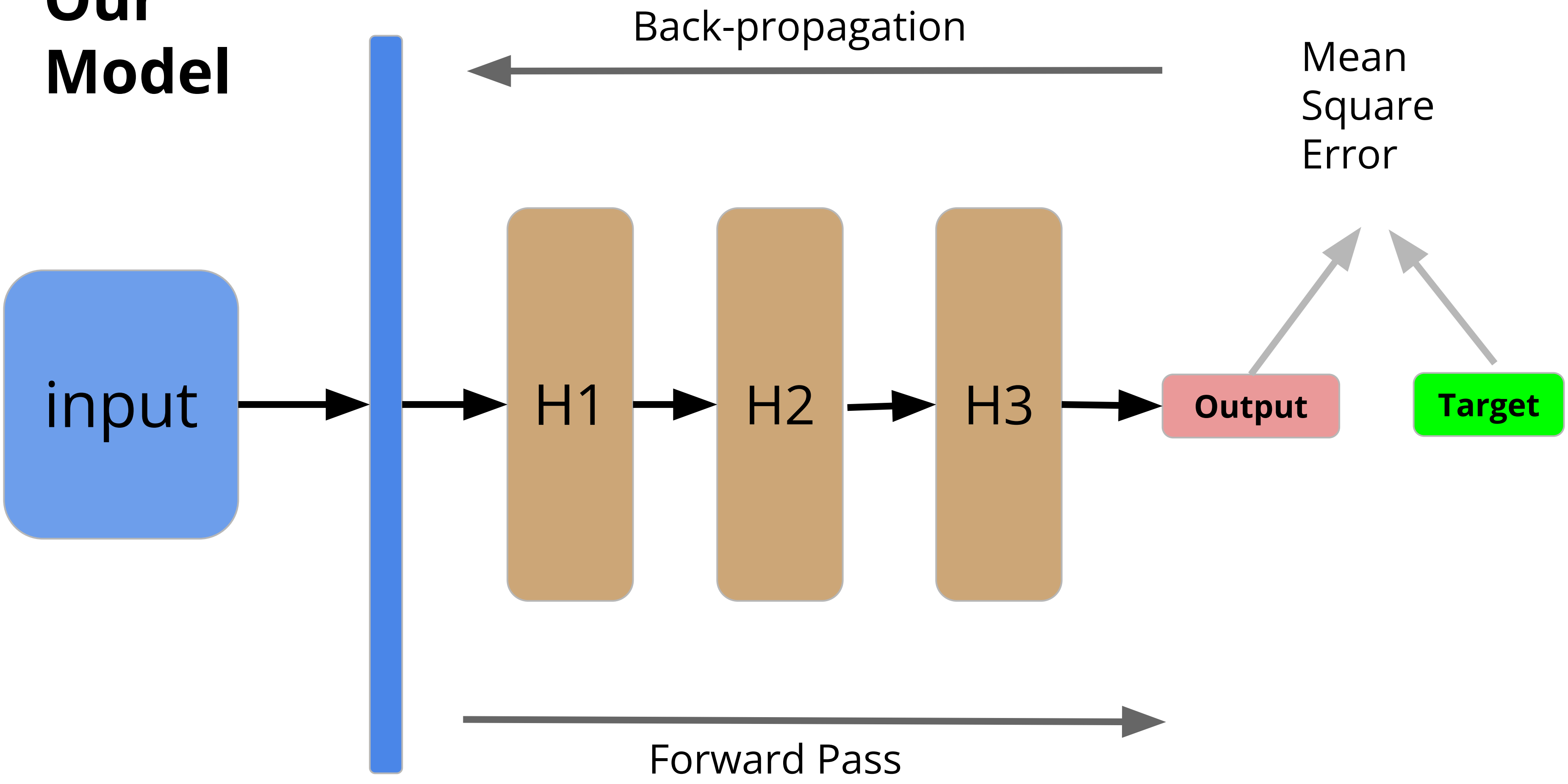


Multilayer Perceptron(MLP)





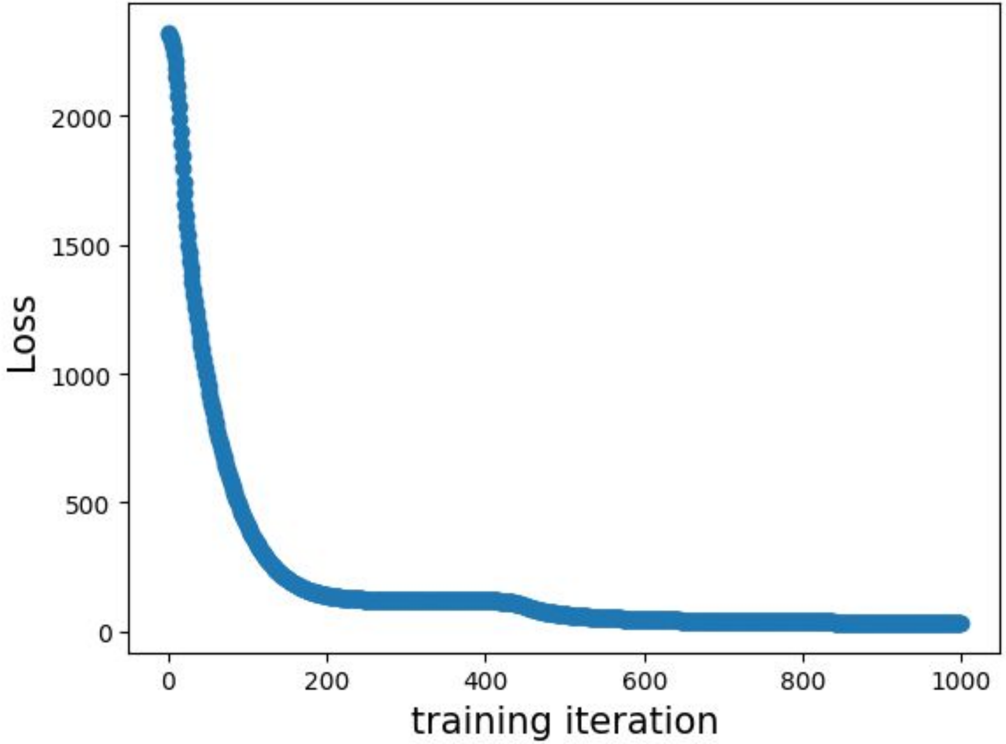
Our Model


Back-propagation
Mean
Square
Error
input
H1
H2
H3
Output
Target
Forward Pass

Learning Performance

Test loss:

135.08



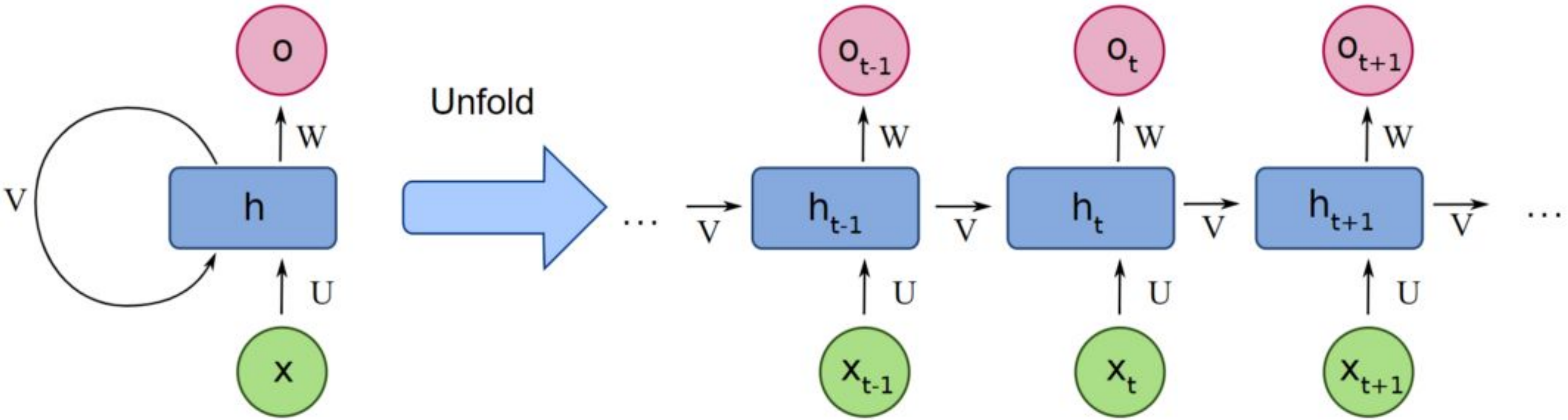
Drawbacks

Flatten the input -> Not fully taking advantage of the sequential nature of the data

Recurrent Neural Network!

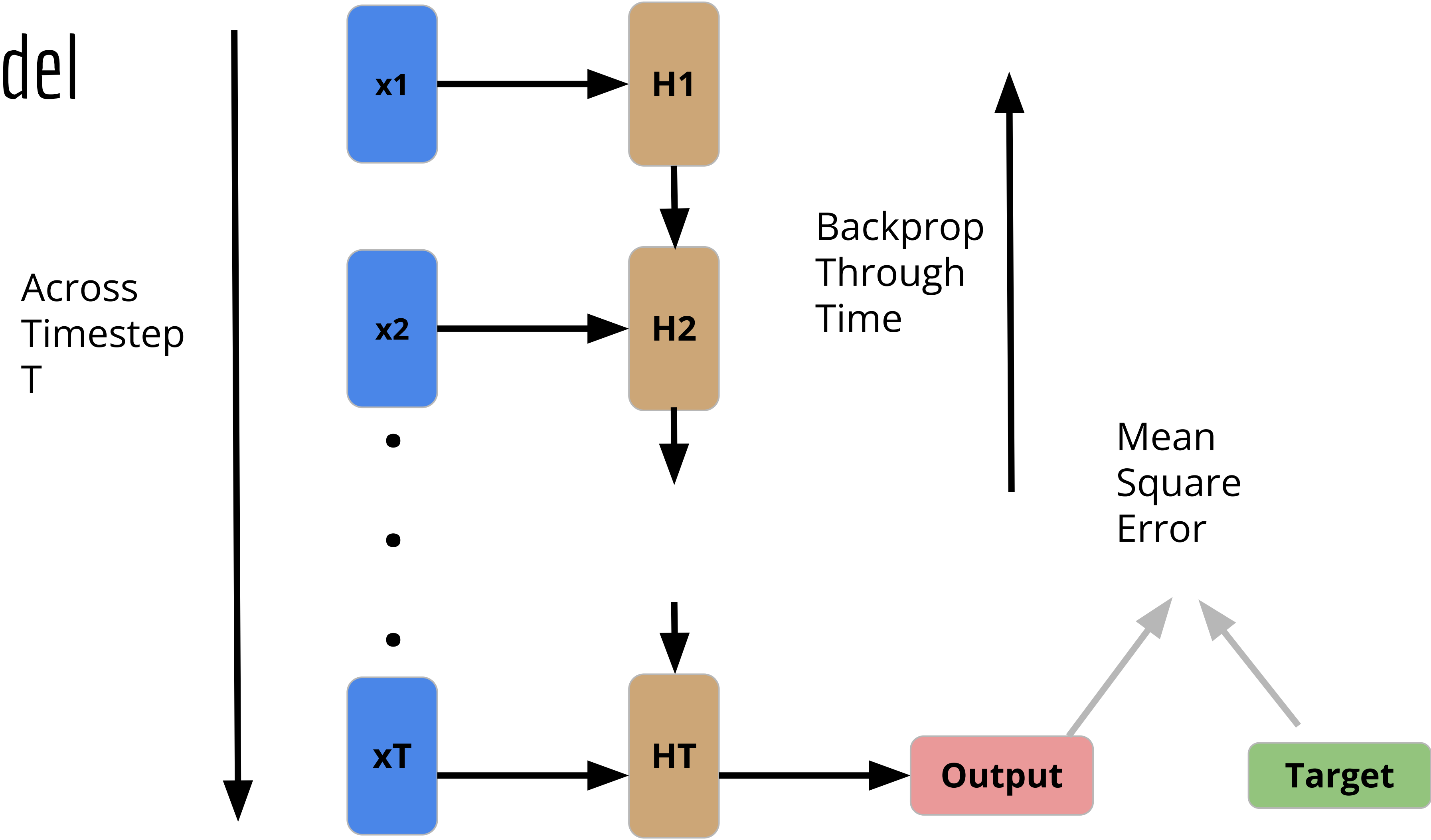
Recurrent Neural Network(RNN)

The input is Sequential! -> an additional dimension for timestep



$$\mathbf{o}_t = \mathbf{W}\mathbf{h}_t \,, \mathbf{h}_t = \varphi(\mathbf{a}_t) \,, \quad \mathbf{a}_t = \mathbf{V}\mathbf{h}_{t-1} + \mathbf{U}\mathbf{x}_t + \mathbf{b}$$

Model

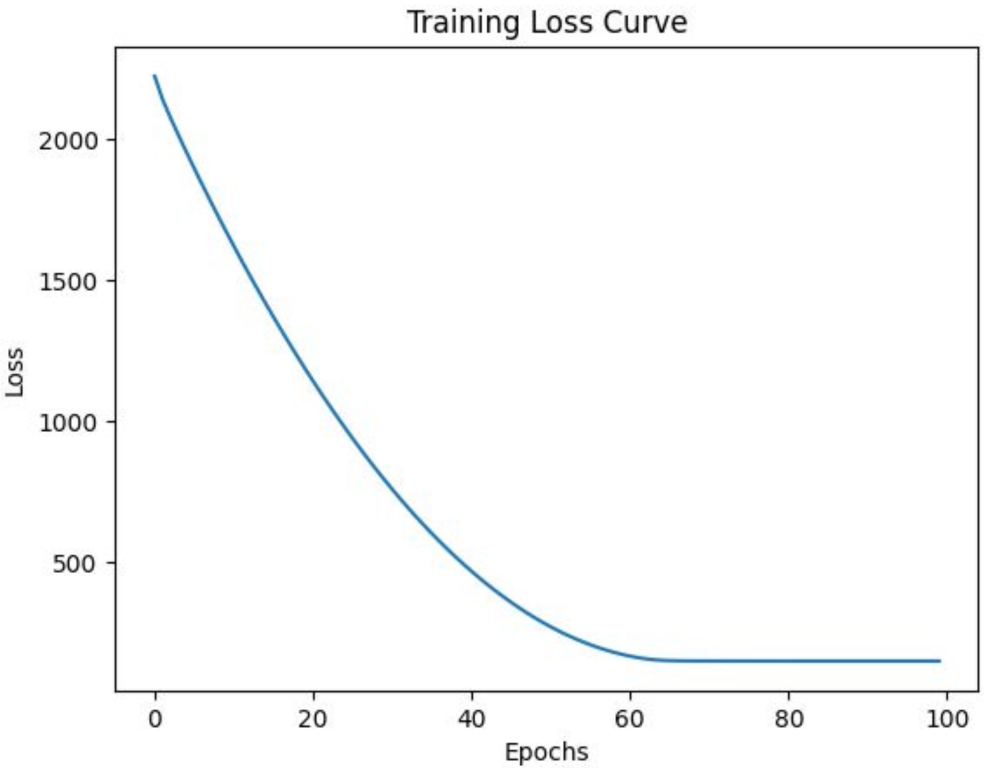


Across
Timestep
T
x1
x2
xT
H1
H2
HT
Backprop
Through
Time
Mean
Square
Error
Output
Target

Learning Performance

Test loss:

143.23



However...

When we test the model on the test set, it always give us the same output regardless of the inputs

Why?

Potential Reason 1:

The number of training data points (629 games) is way less than the sequence length/total number of time steps (nearly 3000).

Potential Reason 2:

Vanishing gradient due to large sequence length and the nature of RNN

Weixuan’s New Approach

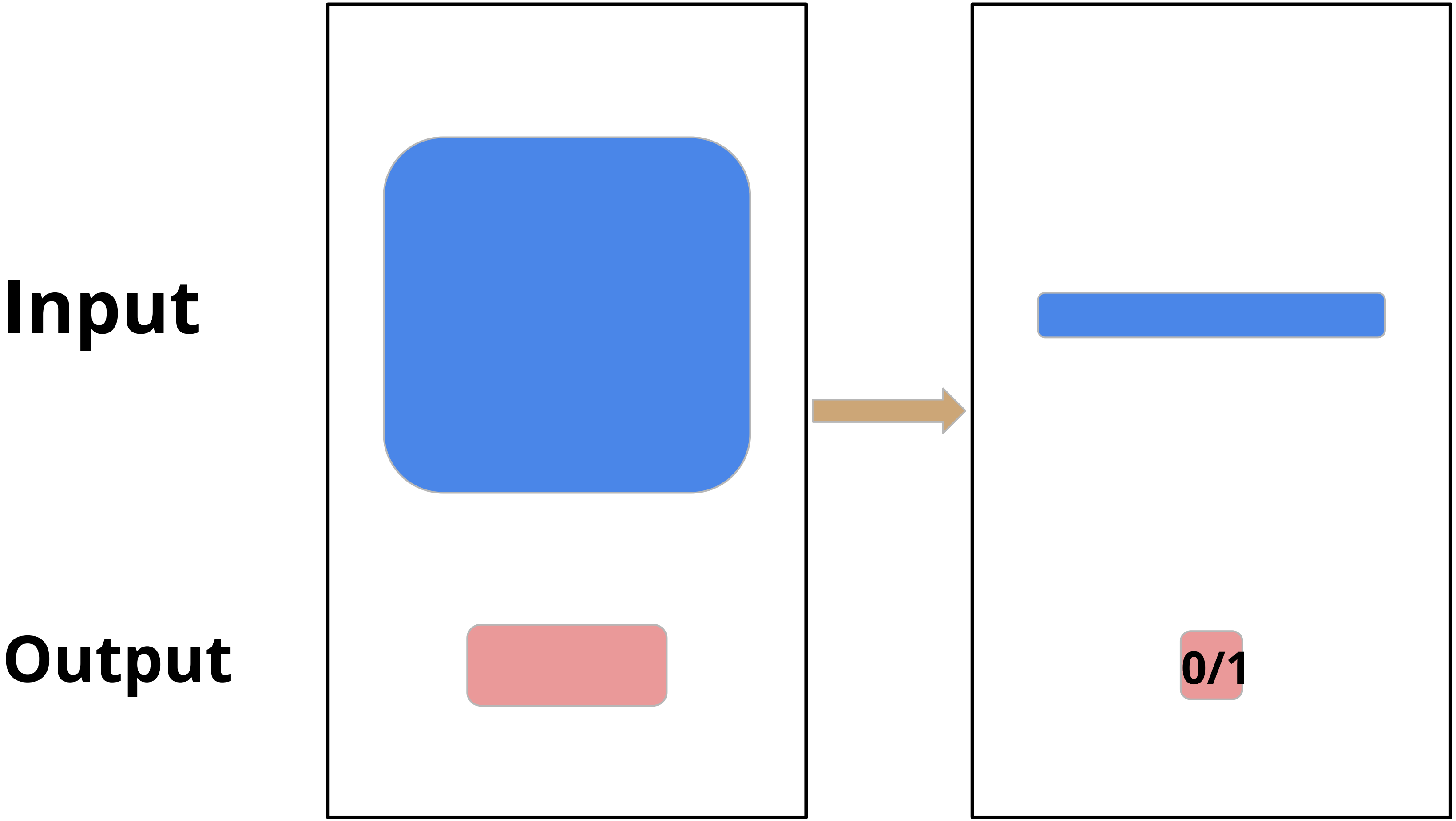
The number of home team’s winning attacks = the total number of attacks that are home team’s winning attacks

New Idea:

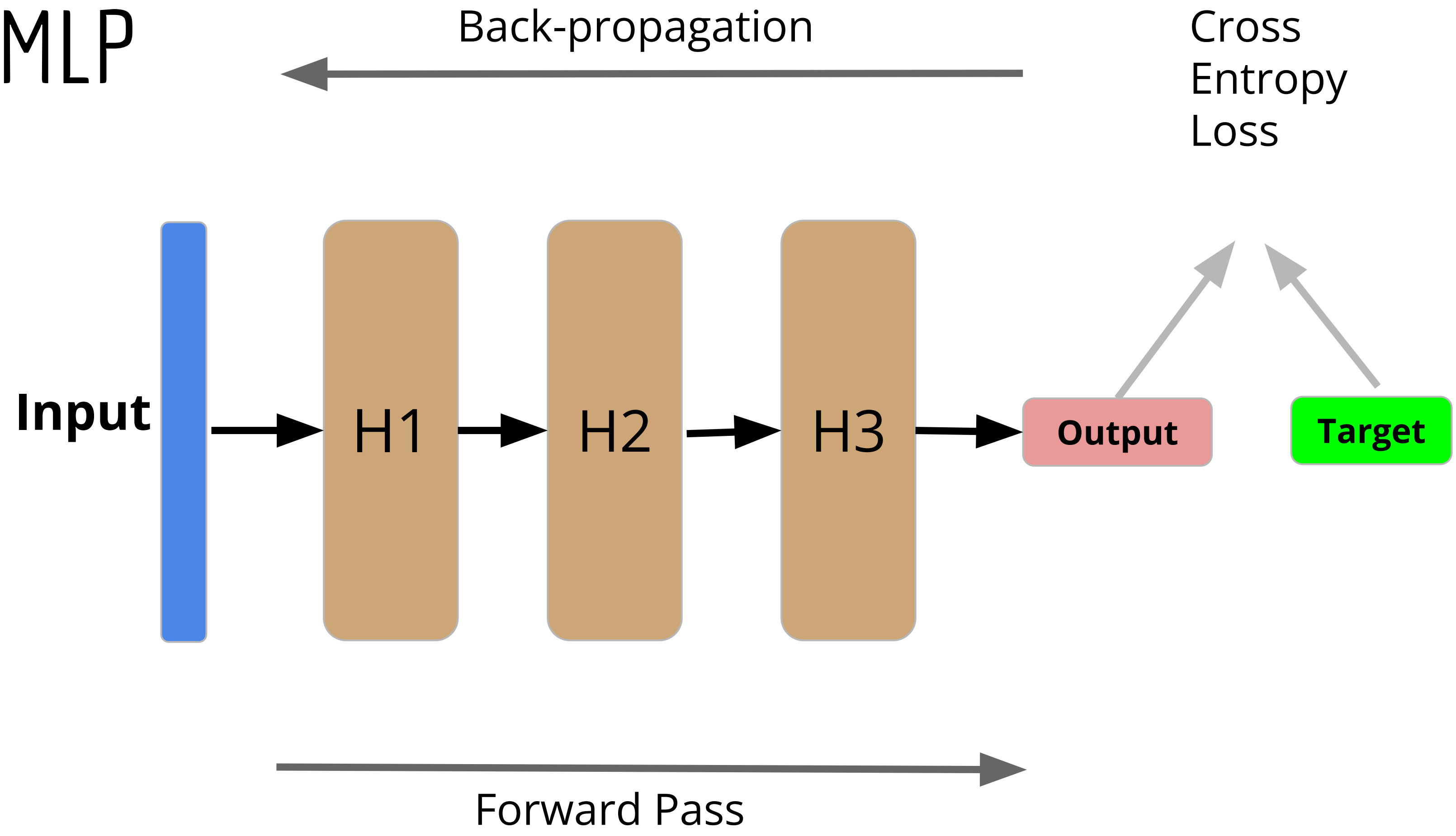
1. Extract the timepoints where an attack happens in one match
2. For each attack, predict whether it is a winning attack of the home team or not: 1- yes; 0 - no
3. Finally sum up all the outputs, which should be the total # of winning attack of the home team

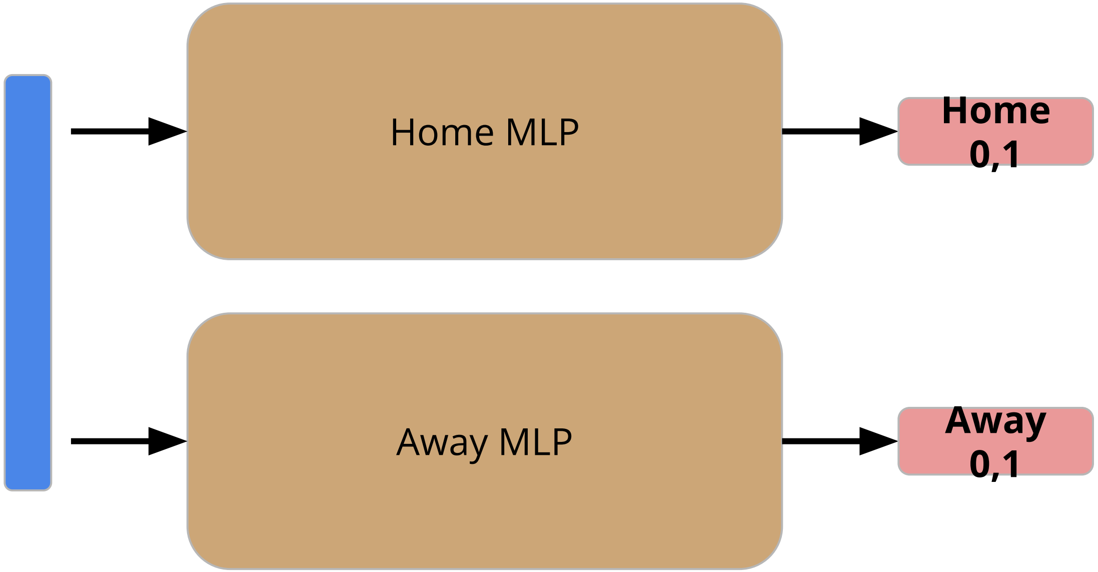
Now 260 data points vs 37 features

Notice that it also breaks the sequential nature of the data


Input
Output
0/1

New MLP


Back-propagation
Cross
Entropy
Loss
Input
H1
H2
H3
Output
Target
Forward Pass



Step 1:
Training Dataset:

First 80% of Attack data

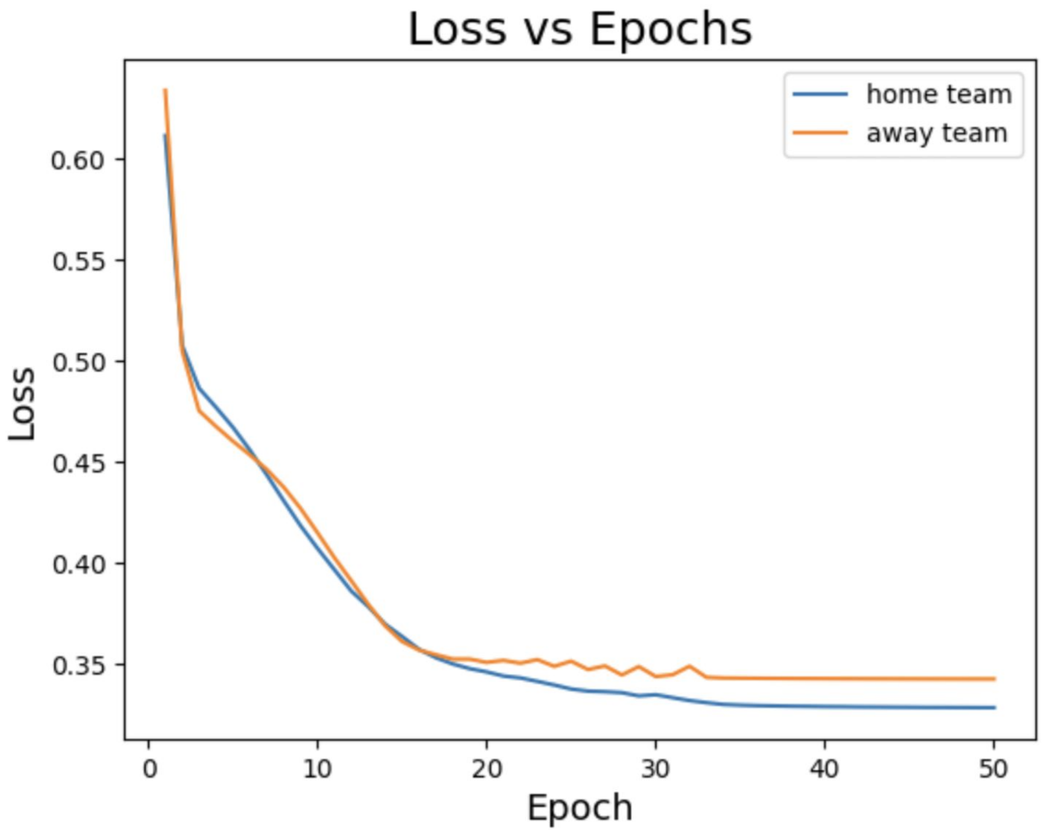
Step 2:
Test the model on the remaining 20% of data

Step 3
Inputs: all attacks data
Outputs: 2 arrays of binary numbers

Step 4:
Sum up all elements in each array

Learning Performance

For one file



Average Across Ten Files

Test Accuracy: 78.84%

Mean Square Loss of the number of winning attacks: 36

Minh’s New Approach

Idea: (Using MLP)

Instead of using the whole game data, for each game, use all sequences that lead to an attack (Dig → Set → Attack).

Inputs: a 2-timepoints corresponding to Dig and Set. (In some cases, where there is only a Dig or a Set, I perform padding)

Output: whether the following attack would be a winning attack or not (0 or 1)

Data Pre-processing

Number of data points: 28901 attacks (from Pac 12 only)

Use One-Hot Encoding instead of Label Encoding.

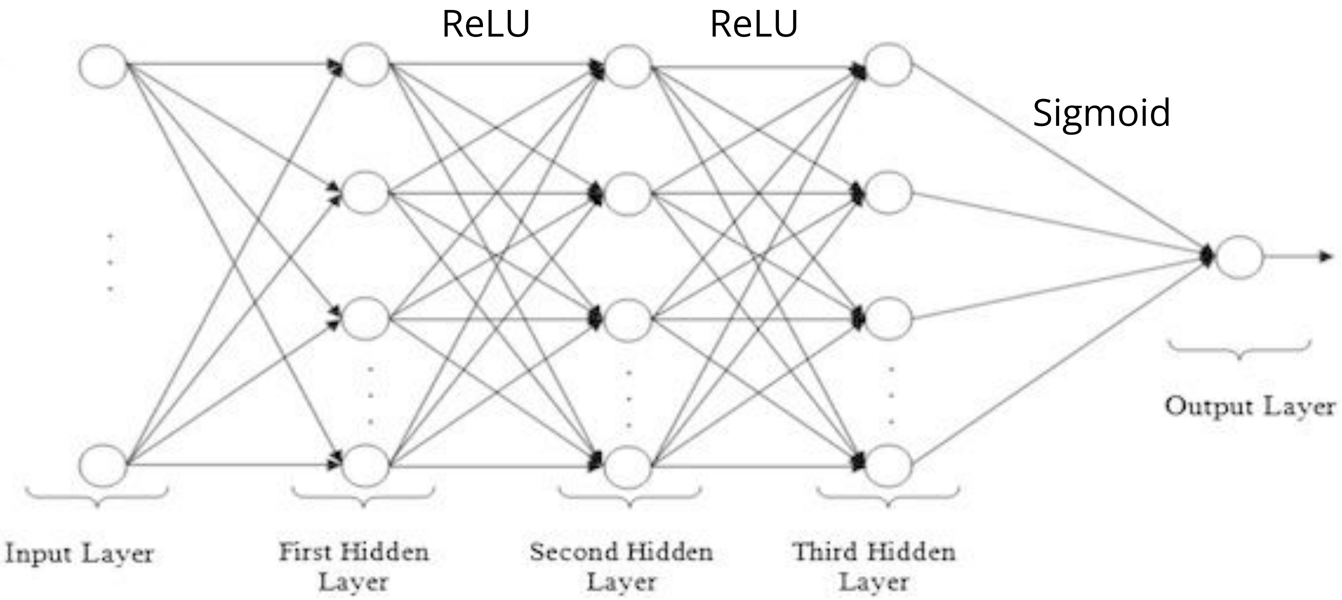
⇒ Now we have 194 features.

Some attacks only have a dig/set preceding them

⇒ Perform padding

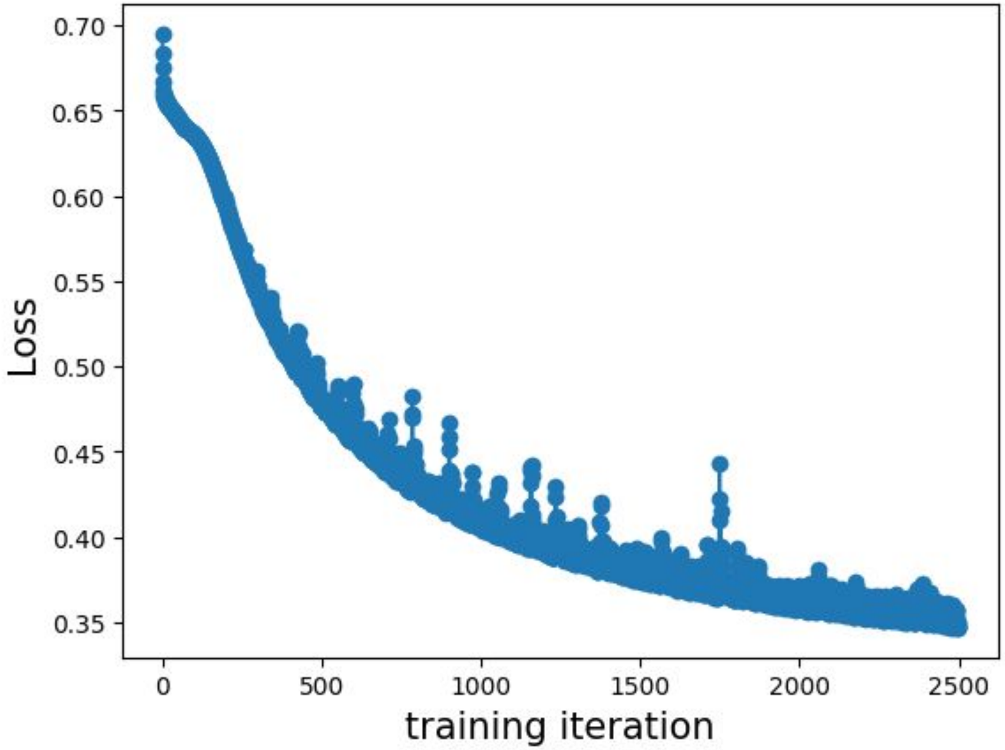
Model

The hidden size is 500.



Learning loss

Note: This only show the first 2500 epochs, I continue training them till 6000 epochs



Result

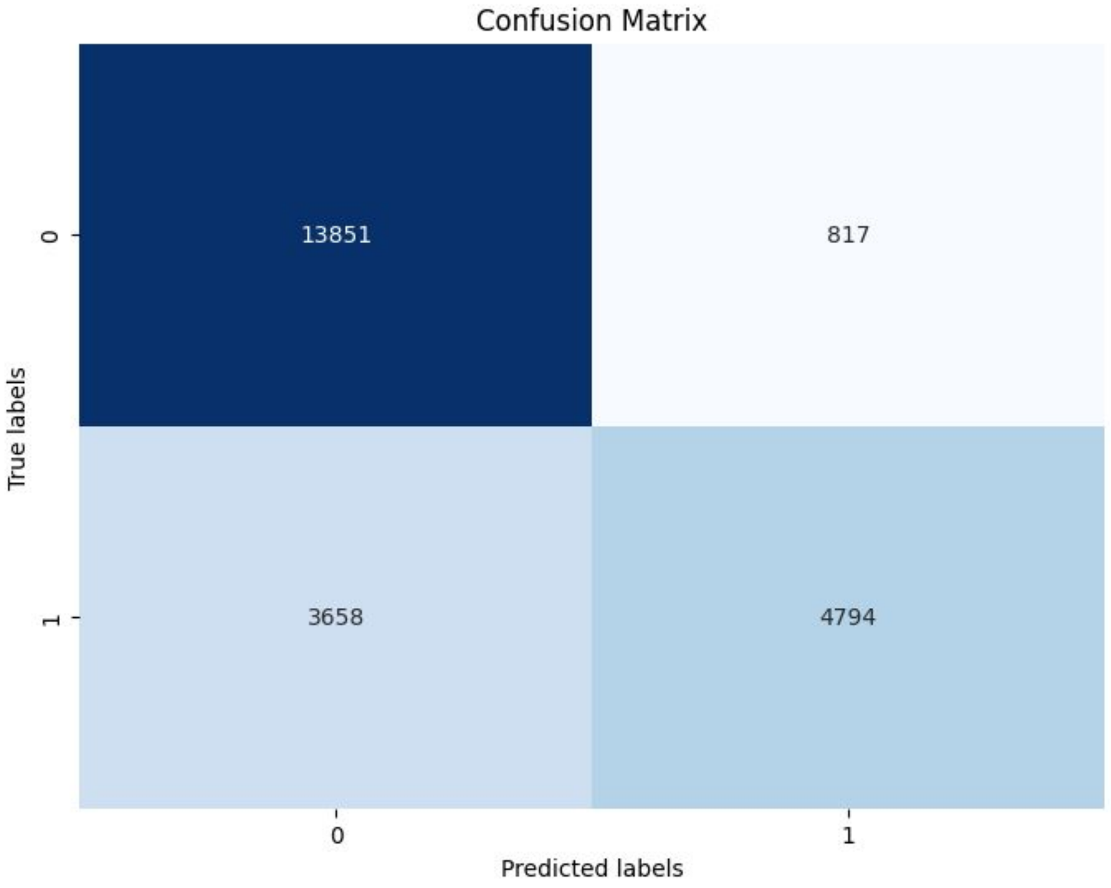
On the test set,

Test Loss: 0.3433

Accuracy: 80.64%

Precision: 85.44%

Recall: 56.72%



How we envision this being used

1. We provide trained model to coach
2. After game, they process video to .dvw file
3. Extract sequences preceding attacks
4. Predict attack outcomes for each sequence → This give expected kills (xK) for a game
5. Compare to actual game results → Analyze whether they under- or over-performed

Future Work

- The current model is only trained on PAC-12 data because of limited computational power. (Google Colab)
- Better data pre-processing for the model
- Determined which features that the model considered as more important, in order to make suggestions for the coach

Our Takeaways from DRP

- Learned new Neural Network Architectures (MLP, RNN)
- Understand the math behind the model we selected (activation function, gradient descend, back propagation,...)
- Better understanding of how to apply deep learning techniques to sports data

Q & A