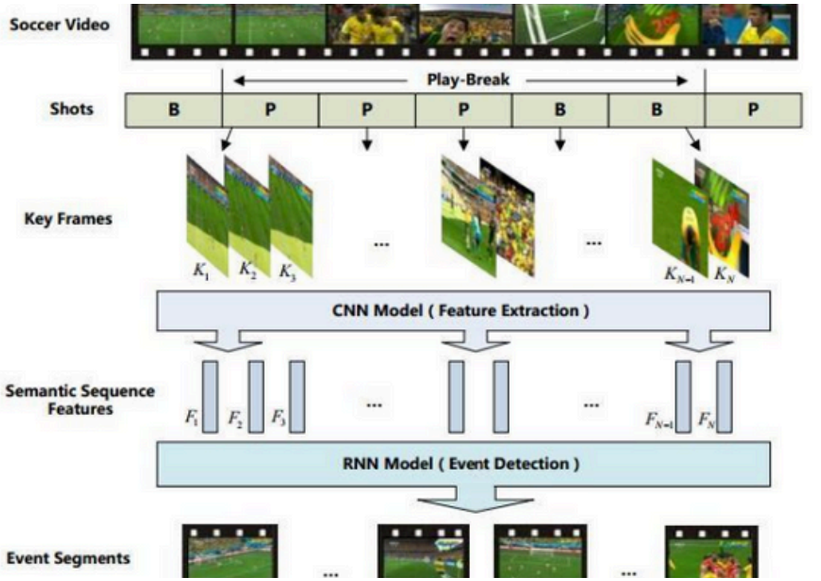
Soccer Foul Detection Based on CNN and Spatial Attention

Xi Chen & QiXiang Jiang

An Analysis of Automatic Soccer Video Event Detection Based On A Deep Neural Network Combined CNN and RNN

Introduction

In recent years, the field of sports video analysis has seen significant advancements, particularly with the integration of deep learning techniques to automate the detection of specific events within game footage.



Model

This paper uses a hybrid network model that leverages the strengths of CNN in feature extraction and RNN in handing time-series data.

The model takes play-break frames as input, and output event prediction as one-hot encoding format.

Performance

Their model achieves 81.25% recall and 92.86% precision on Card prediction.

For other classes, the performance is better on average.

Predicted / ctual	Corner	Goal	Goal Attempt	Card	Missed
Corner	48	3	1	0	2
Goal	2	45	2	0	0
oal Attempt	2	3	66	1	2
Card	0	0	1	13	2
Recall(%)	88.88	91.84	89.19	81.25	-
recision(%)	94.11	88.23	94.29	92.86	-

Our thoughts on the paper

RNN is not needed

From our perspective, Foul is less time-sensitive and space-sensitive compare to other types of soccer event. A foul might happen any time, any where, and is observable in a short period.

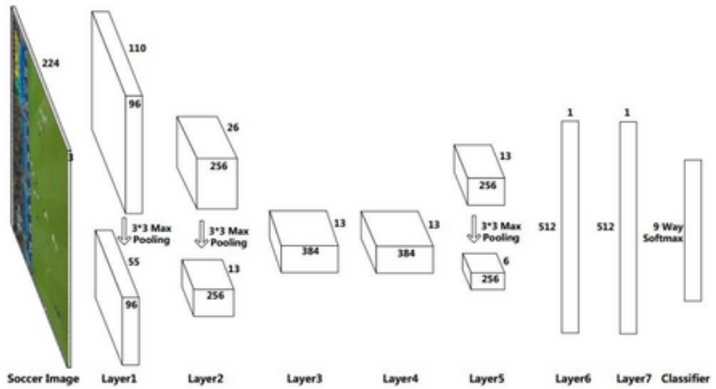
Needs more data

The paper only uses 16 samples for Card class. It is not convincing with such a data size. We should find a much larger data to test the performance of CNN.

Add new technique to model

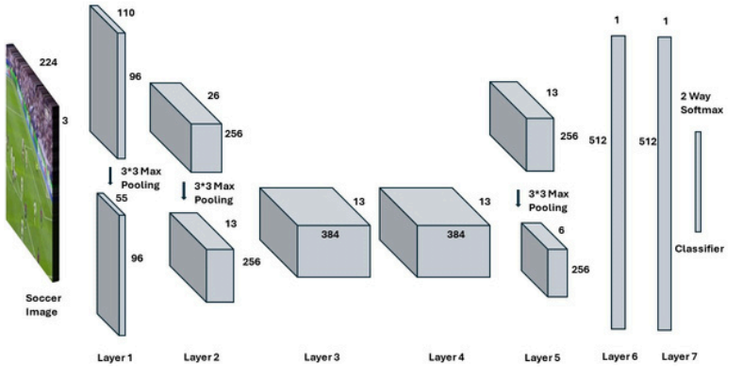
We could add layers to the model to boost robustness. For instance, we could add a spatial attention layer after the last convolutional layer to ensure our model focusing on more important areas of the input. In addition, we need to use GradCam (Gradient-weighted Class Activation Mapping) to see where the model is looking at.

Model Structure Comparison



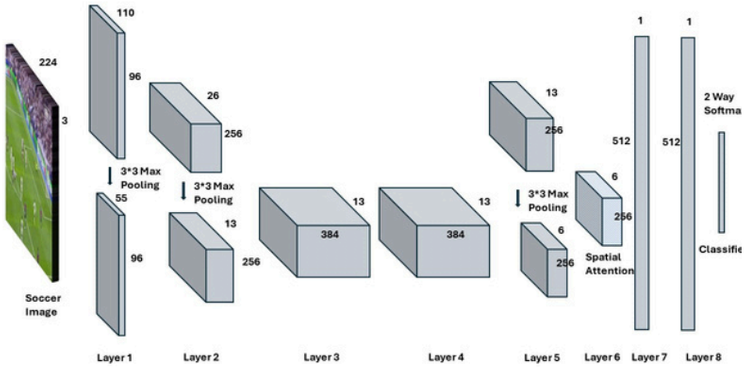
Paper's Model

Consists of five convolutional layers, two fully connected layer and the last output layer that is a 9 way softmax classifier



Normal Model

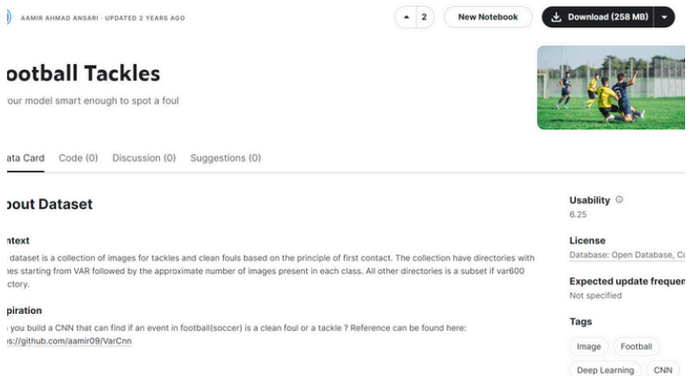
Everything same as the paper's model instead the output layer now is 2 way softmax classifier (foul or no foul)



Spatial Attention Model

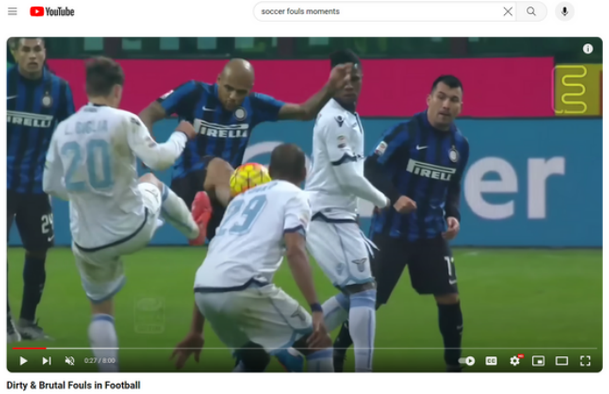
Everything same as the Normal Model except a spatial attention layer is added right after the last convolutional layer

Data Collection



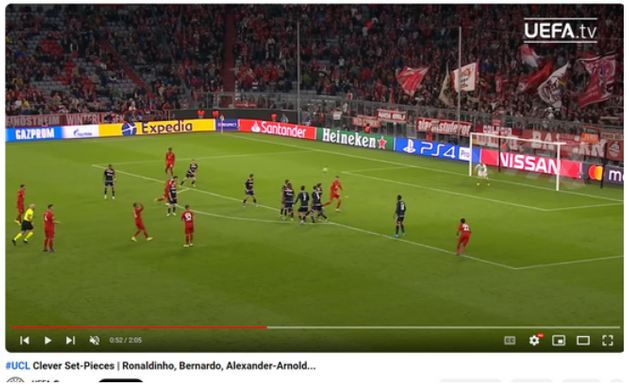
Kaggle Tackles Dataset

Contains ~600 clean tackles and ~600 fouls image data



Youtube Screenshot

In order to increase fouls data, we search “fouls moment” on youtube and take screenshot when they foul. (~200 data)

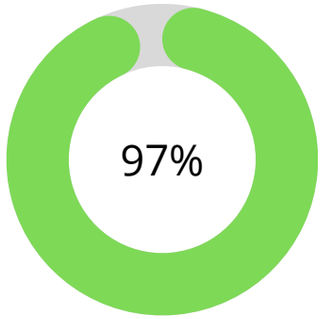
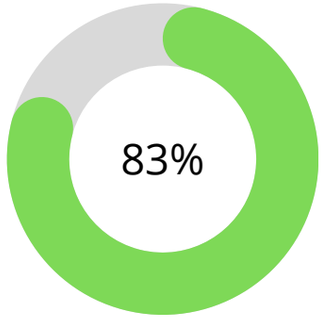
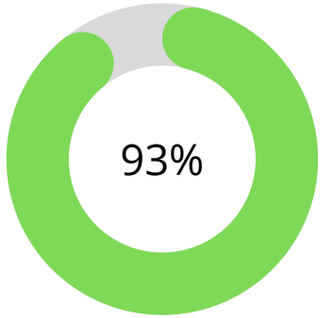
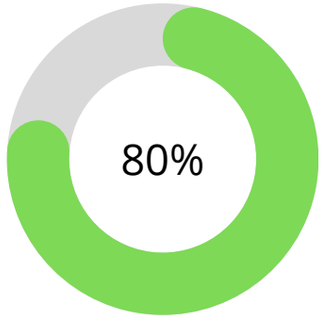


UEFA match highlights

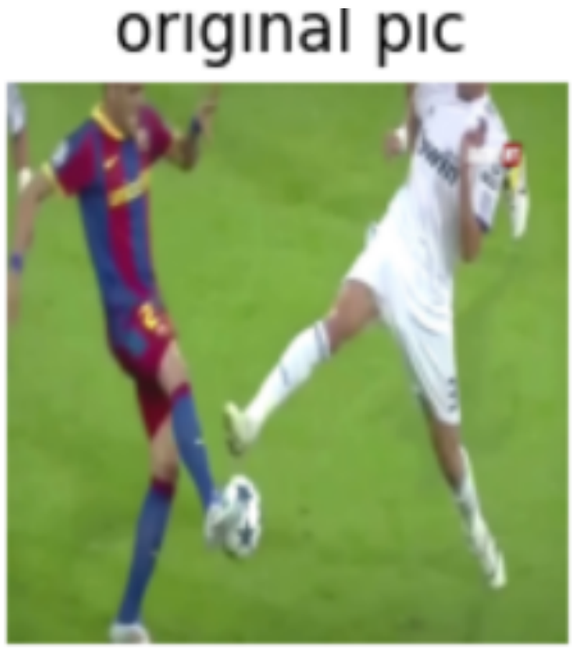
UEFA (Union of European Football Associations) uploads highlights every week, which are high quality data for no foul data. (~2000+ data)

Train Params & Performance

We uses SGD as our optimizer with learning rate of 0.001, momentum of 0.9. We choose CrossEntropy as our loss function. The number of epochs is 50.

	Accuracy On Train Set	Accuracy On Test Set
Normal Model		
Attention Model		

Prediction on the same frame

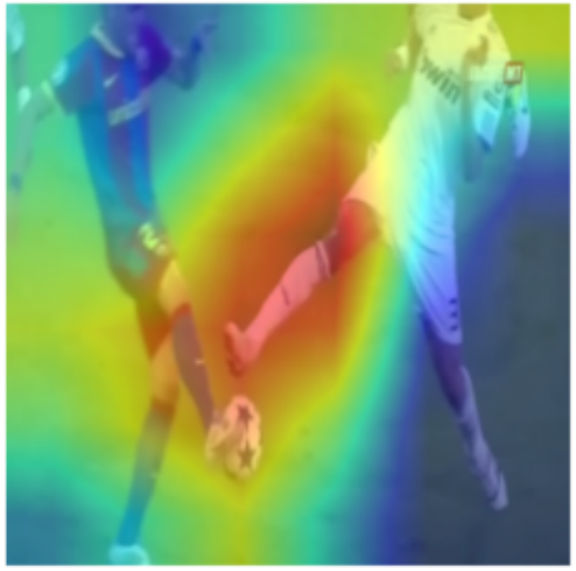


The Original data

watch your feet!

GradMap for normal model

the model mainly focuses on left shoulder and feet of Pepe



GradMap for attention model

the model focuses on the right leg of Pepe, right leg of Alves and the ball


Time To Try It Out!
Emirates
FLY BETTER

Conclusion & Future Work

Attention Model makes more sense

Although the normal model achieves higher accuracy than the attention model, the GradMap shows that attention model uses evidence that is more relevant. It shows the potential to become more sophisticated deep neural network to predict fouls.

More data is needed

We increased the size of our dataset, but we forgot to ensure the distribution of different type of data (muti-view) is even across our dataset. Diversity matters!

Contextual information needed?

Since we are only predicting one frame at a time, the result could be bias. One paper takes Live Action, Replay 1, Replay 2 as input and produce prediction for their combination (7 in total), which we could refer to.