Stanford University

SOCCER STATS WITH COMPUTER VISION

Hayk Tepanyan {tehayk@stanford.edu}
Department of Computer Science, Stanford

Problem Formulation

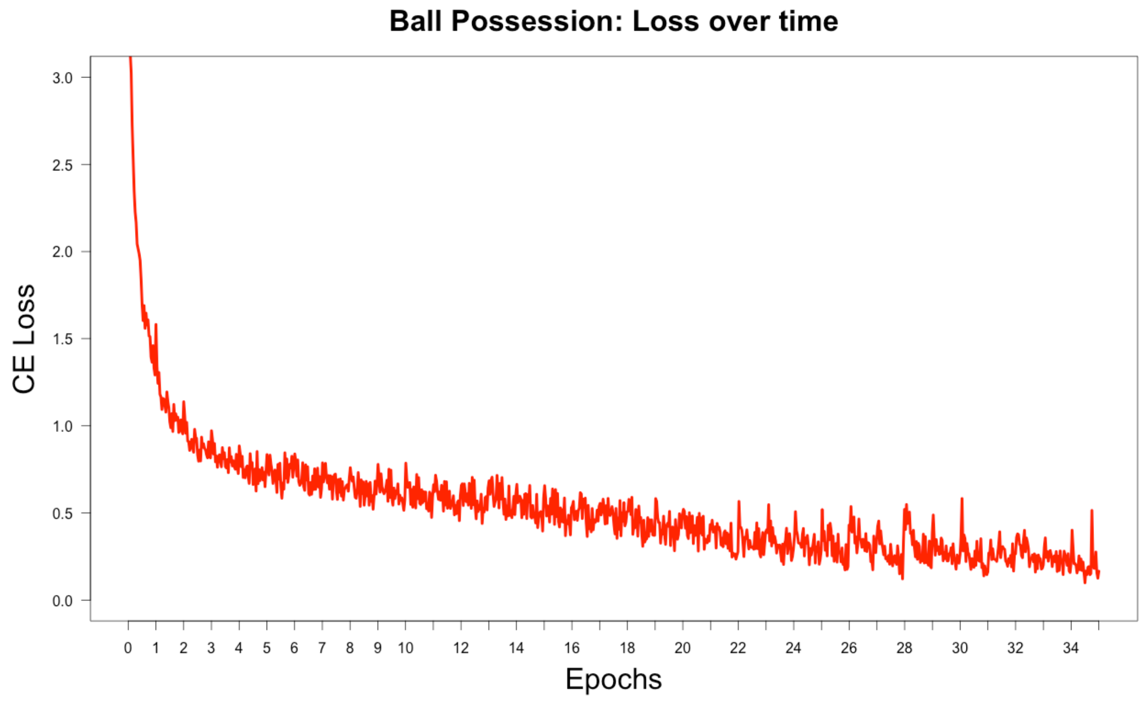
We tackle two specific problems: finding out the ball possession and localizing the ball. The first problem is very challenging even for humans. Currently leading sports agencies determine ball possession using number of passes by each team. While this number is highly correlated with the actual ball possession time, it is skewed for teams like Barcelona and Arsenal that tend to make more frequent passes on average. We approach these two problems separately.

Dataset And Details

We use frames from a single match. We manually labelled 4000 frames out of 140K frames from a 90 minute recording of Real Madrid vs Barcelona game. For each frame we labelled the ball possession and location. To capture diverse frames and not lose temporal info at the same time, we labelled 80 episodes of 10 seconds, each containing 50 frames, capturing every 5th frame.

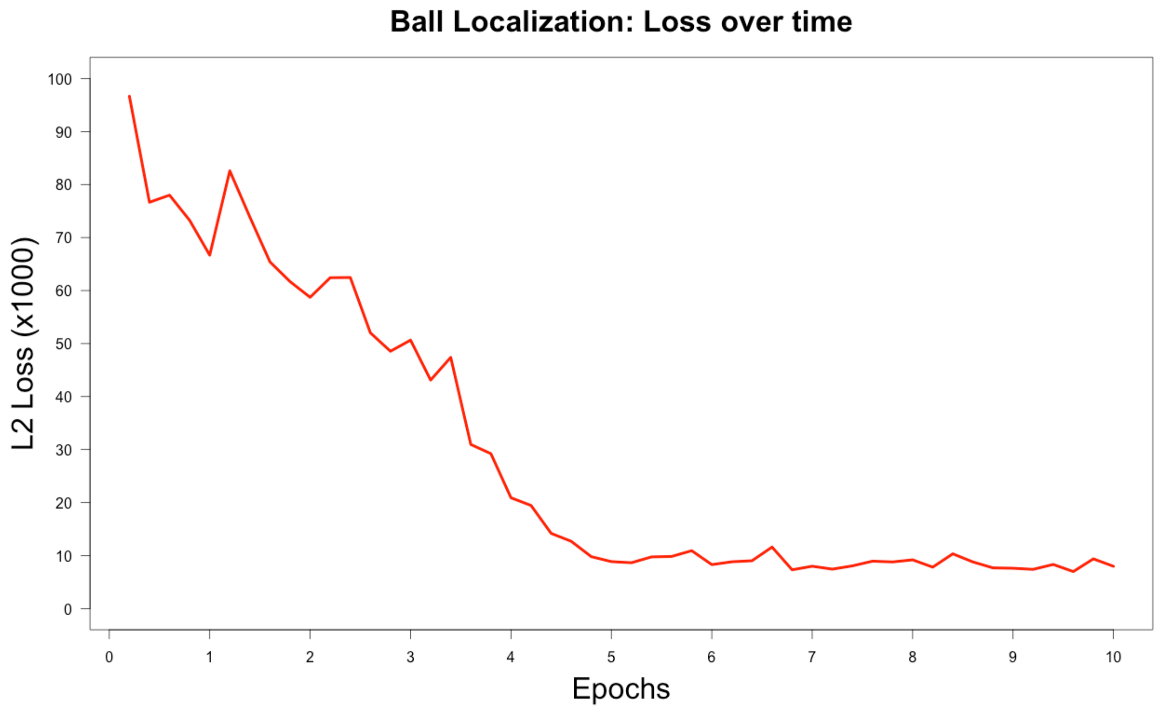
Solving Ball Possession

For the team possession task we use a vgg-style CNN. More specifically the network consists of 7 (3,3) 2D-Convolution layers followed by max poolings, ending with 5 fully connected layers. There is a dropout layer after the first FC layer. We use cross-entropy loss with L2 regularization for weights in large FC layers. We use 90% of the data as training, 5% as validation and 5% as test data. We measure the accuracy as fraction of correctly guessed frames. For each frame we guess the category it belongs to: Team 1, Team 2, Game-Off. The test accuracy for the best model is 85.5%

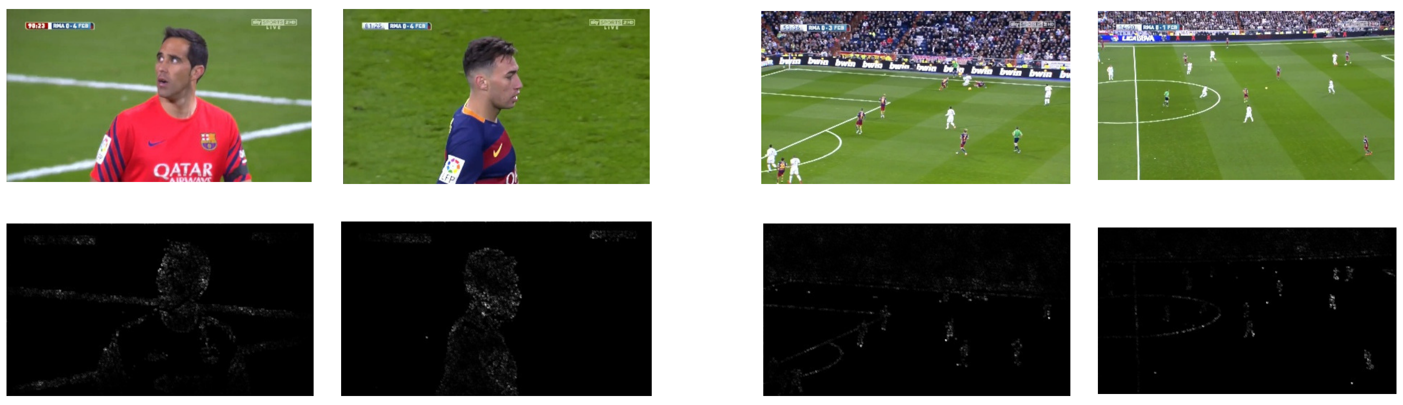


Solving Ball Localization

The high level architecture for the ball localization task is the following: (3D-Convultion -> pool) x N -> RNN -> (FC with ReLU) x M. The input is a list of 5 successive frames within an episode. The max poolings do not reduce the dimension of the depth, so after the CNN layers we end-up with features for each frame in the input that we can feed in to the RNN. We predict the (x,y) pair for the last frame in the list and use L2 loss. We measure the accuracy of the model by the average distance between the real and predicted ball positions. The test accuracy is 58px.



Saliency Maps



By comparing the saliency maps, we can see how the network distinguishes between the game-on and game-off frames. On the left we have 2 game-off frames, where the edges of the large silhouettes are highlighted. On the right, in the game-on frames the activations are around the smaller silhouettes of players.

Insights

The high performance on team possession prediction suggest it might be possible to build a universal model not for a particular game, but for any soccer match. The fact that our model actually calculates the ball possession for each frame instead of using number of passes gives hopes that one day such a technique might substitute current way of gathering soccer stats.

Whereas the team possession task was easy for the CNN to learn, the ball localization task is much harder. The reason we suspect is the scarcity of the labelled data and the fact that the ball covers a tiny fraction of the image, and sometimes is not visible at all. Working with a larger dataset of higher quality images may solve this problem.