Some notes regarding format for the soccer data:

Included in the data set are a training set with 7500 sequences and two separate set of sequences for testing. Python users can load the training data (in pickle format) into a dictionary, the key for each sequence would be "***sequence_n***",where n is anywhere from 1 to 7500.

Each sequence contains a segment of tracking data corresponding to actual game play from a recent professional soccer league. The format of each sequence is as follows:

- Each sequence is a matrix (numpy 2D array) with 46 columns. Each row contains 23 pairs of (x,y) coordinates of 22 players from both teams and the ball at frequency of 10Hz.
- In the first 22 columns are 11 (x,y) pairs of defense team. The following 22 columns are coordinates of attacking team (defined as the team with consecutive possession of the ball). The last 2 columns are coordinates of the ball.
- Each set of 22 columns for both attacking and defending team consist of (x,y) pair for the goalkeeper, followed by 10 consecutive (x,y) pairs for the other 10 teammates. The identities and teams vary from sequence to sequence. However, within each sequence, the identity is consistent. Thus concretely, out of the 46 columns from each sequence, we know that the first 2 columns represent the coordinate of defense team's keeper. Columns 2 to 22 contain 10 consecutive (x,y) pairs of other defensive players. Columns 23 and 24 carry x and y coordinates of the attacking team's keeper. Columns 25 to 44 contain 10 consecutive (x,y) pairs of other attacking players. Columns 45 and 46 carry x and y coordinates of the ball.
- The coordinates generally belong to the [-52.5 meter, +52.5 meter] range along the x-axis, and [-34 meter, +34 meter] range along the y-axis, with the very center of the pitch being [0,0]. So for example, to normalize the data to the range [-1,+1], one can simply divide the x-columns by 52.5 and y-columns by 34 (this effectively will re-scale the pitch, which roughly corresponds to soccer field of size 105mx70m, from a rectangular box to a square box)
- The coordinates were also adjusted so that the attacking team will moves from left to right, meaning the defending team defends the goal on the right hand side.
- In aggregate, the data set amounts to equivalently and approximately 45 games worth of playing time, with redundant and "dead" situations removed.