Play-by-Play Sports Modeling with Deep and Recurrent Neural Networks

Jake Moorhead, Josh Osborne, Michael Remington, Sam Kaplan and Brian Hutchinson

Computer Science Department, Western Washington University

Overview

Motivation: Predicting play and game outcomes of sporting events at a play-by-play level has the potential to tremendously improve in-game decision making and roster construction.

Goal: Estimate accurate play and game outcome probabilities and gain insight into the role of individual players in these outcomes.

Approach: Professional and collegiate sporting events can be viewed as complicated dynamical systems, which begin in some initial state and evolve as the event progresses. We model these dynamics using deep and recurrent neural network architectures, in order to predict play and game outcomes.

Approach

We learn two prediction functions:

- Play outcome

$$f_p : \mathbb{X}_1 \times \mathbb{X}_2 \times \cdots \times \mathbb{X}_i \rightarrow \mathbb{P}$$

- Game outcome

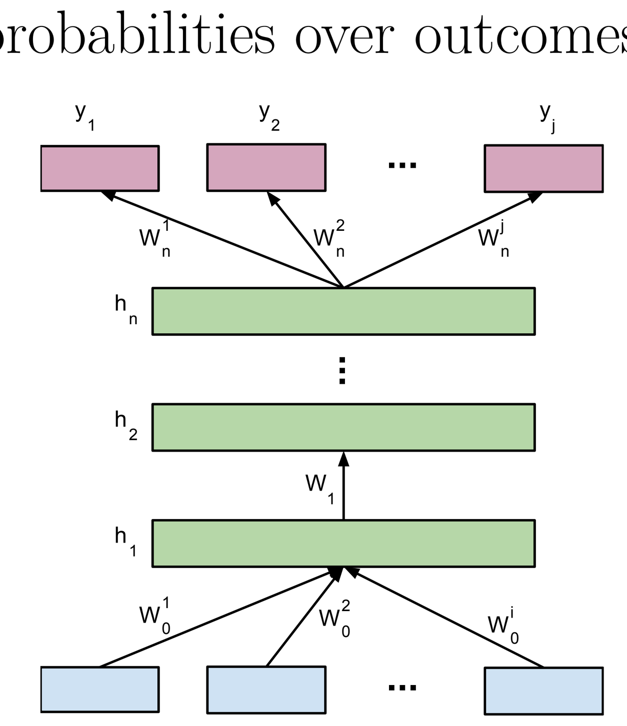
$$f_w : \mathbb{X}_1 \times \mathbb{X}_2 \times \cdots \times \mathbb{X}_i \rightarrow \mathbb{W}$$

- $\mathbb{X}_i$ denotes space of arbitrary input features
  - e.g. number of outs, balls, strikes
- $\mathbb{P}$ denotes the space of baseball play outcomes
  - e.g. single, double, strike-out
- $\mathbb{W}$ is the set of possible game winners
  - home or away

Models

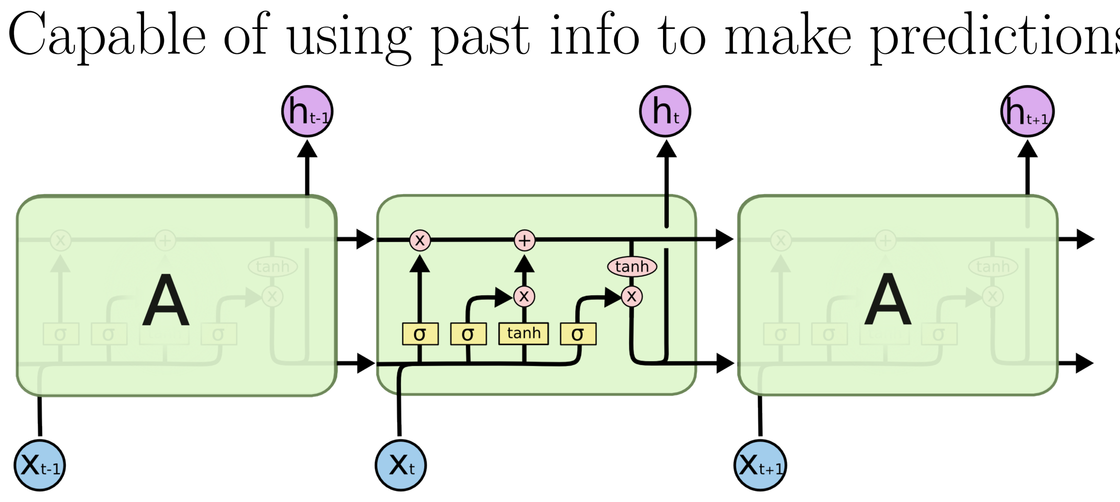
Deep Neural Networks

- Provided with game state data as input
- Learn multiple non-linear transformations
- Output probabilities over outcomes



Long Short Term Memory Networks

- Maintains hidden rep. of current & past data
- Capable of using past info to make predictions



http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Experimental Setup

Data

- 47 input features
- Mix of categorical (e.g. players, teams) and continuous (e.g. strikes, balls) data
- Categorical data rep. by one-hot vectors:

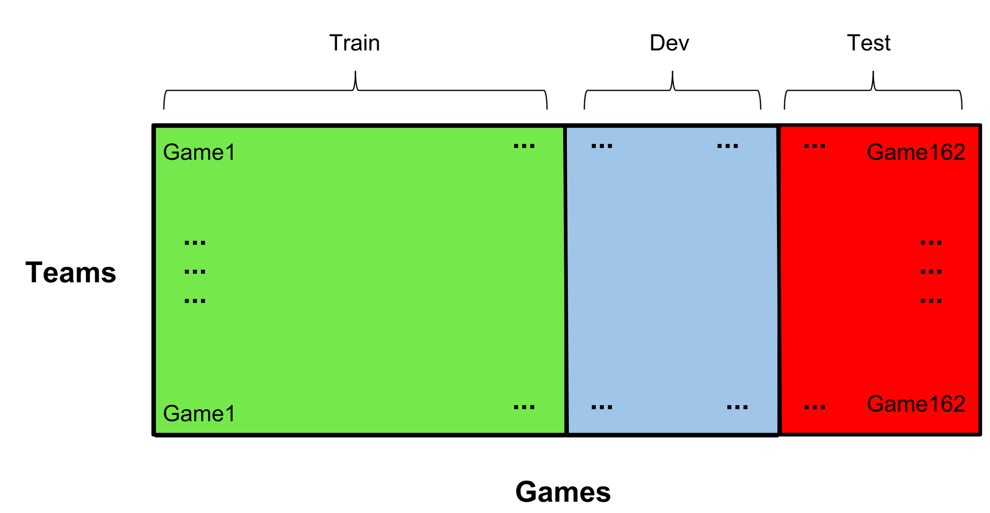
$$[0 \ \cdots \ 0 \ 1 \ 0 \cdots 0]^T$$

- Continuous data are $x \in \mathbb{R}$ and normalized by:

$$\frac{x - mean(\vec{X})}{\sigma(\vec{X})}$$

- Data split into train, dev and test

Set	Plays	Games
Train	133,578	1701
Dev	28,063	364
Test	28,341	365



Results

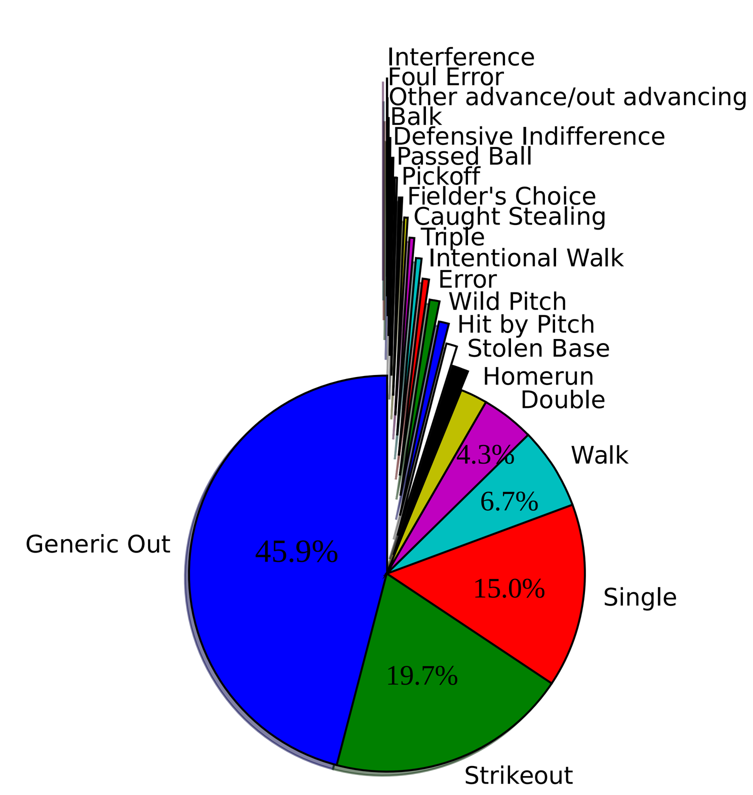
Model	Accuracy
DNN	53.4%
LSTM	52.4%
Baseline	45.9%

(a) Play Outcome

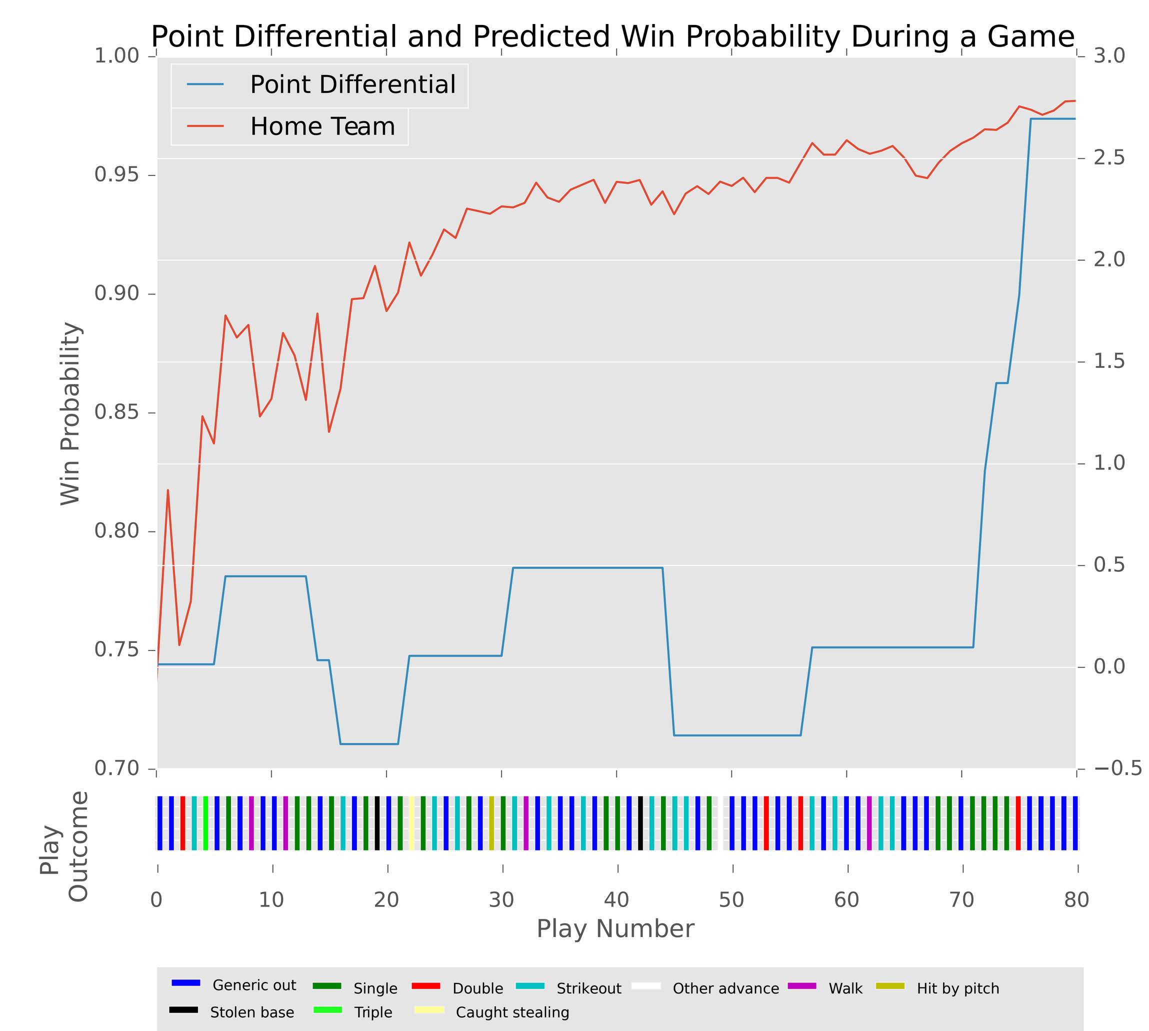
Model	Accuracy
DNN	77.3%
LSTM	77.9%
Baseline	76.1%

(b) Game Outcome

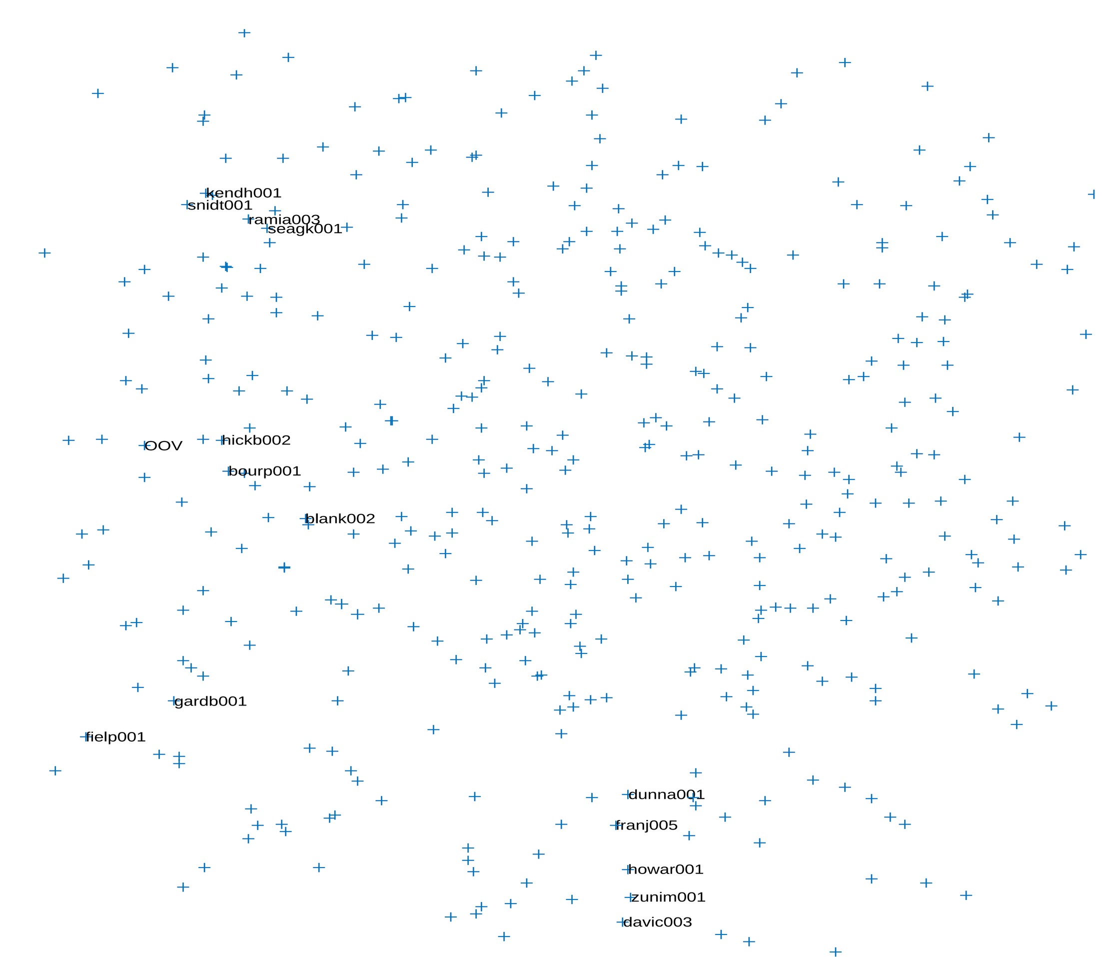
- Relative reduction of 14% and 8%
- Accuracy may be too coarse as a metric



Analysis



- Trends beyond score affect win expectancy
- High dimensional representations of players



- Similar players are close to each other

Conclusions & Future Work

- Basic play state information (outs, strikes, balls) most important for play outcomes
  $\Rightarrow$ Pitcher and batter also important
- Switch evaluation metric to perplexity (reward for good probabilities)
- Apply to other sports (e.g. football)

Acknowledgements

The authors thank

- Joe Renner for his valuable insight and analysis in the interpretation of the results, and
- the Nvidia Corporation for their donation of the the Titan X GPU used in this research