Deep Reinforcement Learning in Ice Hockey for Context-Aware Action Values and Player Ranking

Paper ID: 3945
Content Area: Applications of Reinforcement Learning

Abstract

Deep reinforcement learning (RL) has achieved impressive applications in virtual games, but none in physical professional sports. In this paper, we describe a new application to a fundamental problem in sports analytics: assessing the performance of players. A common approach is to rank players by the quality of the actions that they perform, which raises the question of how to quantify the impact of their actions. The action-value concept from RL provides a powerful AI-based answer to this question, where action values represent expected returns. Specifically in our hockey model, action values represent the chance of scoring the next goal. Using deep RL techniques, an action-value Q function is learned from a massive dataset with over 3M play-by-play events in the National Hockey League. To capture dependencies among recent events while looking ahead to the long-term effect of an action, we introduce a novel LSTM architecture based on puck possession. The trained network provides a rich source of knowledge about how the value of an action depends on the match context, the rink location, and the game time. We show how the learned action values can be used to quantify the impact of player actions on goal scoring, and to rank players by their cumulative goal impact. This ranking identifies undervalued players, is consistent throughout a play season, and correlates highly with independent standard player metrics (e.g. $\rho > 0.93$ for total points (= goals + assists).)

Introduction

A fundamental goal of sports statistics is to quantify how much physical player actions contribute to winning in what situation. As more and larger "play-by-play" datasets for sports events become available, there is increasing opportunity for large-scale machine learning to model complex sports dynamics. The research described in this paper applies deep reinforcement learning (RL) on a massive dataset about player actions, comprising over 3M game events along with x-y rink locations and continuous time stamps. The dataset contains 2015-2016 regular season games in the National Hockey League (NHL). We learn an *action-value or Q* function that represents the chance that a team scores the next goal, given the current match context. The trained network supports several applications in sports analytics; our target application in this paper is the "moneyball" problem of ranking players.

Motivation. A neural net Q-function representation offers several advantages for sports dynamics.

Knowledge Discovery. A Q-function represents the likely consequence of an action in a given game context. It can be used to assess the value of different tactics, which are often disputed by hockey experts and fans (Cervone et al. 2014a).

Neural Representation. A neural network representation, which easily incorporates continuous quantities like rink location and game time. The neural net *generalizes* value estimation to unknown game states and can therefore evaluate new unexplored hockey strategies. We apply model-free Temporal Difference (TD) learning, which does not require explicitly modelling state transition probabilities. This facilitates learning in a continuous state space, without assuming that transitions satisfy a particular parametric form or spatio-temporal stationarity.

Action Values and Player Ranking. A common approach to player ranking is to use *total action values*: assign values to actions and then rank players by the value sum of their actions (Schuckers and Curro 2013; McHale, Scarf, and Folker 2012). For example, the NHL Plus Minus(+/-) score (Macdonald 2011; Spagnola 2013) assigns +1(-1) to player on ice when their team (opponent's team) scores. The Q-function offers two key advantages for assigning values to actions (Schulte et al. 2017; Decroos et al. 2017b): 1) Looking ahead to expected future outcomes scores *all* actions on the same scale, including defensive actions (e.g. checks, blocks), rather than scoring only offensive actions immediately relevant to goals (e.g. shots). 2) Action values reflect the match context in which they occur. For example, a late check near the opponent's goal generates different scoring chances than a check at other locations and times.

Deep Learning Approach. Unlike most previous work on RL which aims to directly compute optimal strategies for complex continuous-flow games (Hausknecht and Stone 2015; Mnih et al. 2015), we solve a prediction (not control) problem in the *on policy* setting (Sutton and Barto 1998). We introduce a possession-based Long Short Term Memory (LSTM) architecture that takes into account the current play history. Our main methodological contribution is to set the key LSTM trace length parameter *dynamically*, so that history traces are confined to a single ice hockey play marked by a team's possession (Dynamic Play Trace DP-LSTM). On our dataset, the training error for our DP-LSTM method decreases smoothly and quickly (below 0.006 after 10 epochs). The trained model fits the data well, in the sense that the predicted scoring chances match the observed scoring chances for a given match context (average KL divergence below 0.015).

Contribution. The main contributions of this paper are: 1) We develop a novel application of deep reinforcement in sports analytics: learning a context-aware Q-function from a massive complex data set comprising 3M ice hockey events. 2) We apply the learned value function to derive a new metric for player performance that utilizes all player observations. To our knowledge, this is the first performance metric that can be interpreted both as a total action value count and a value-above-replacement score. 3) We extend LSTM with a novel possession-based method for setting trace length that is appropriate for sports dynamics.

Related Work

We discuss the previous work most related to our approach.

Deep Reinforcement Learning. Among the previous deep RL work on control in continuous-flow games, (Hausknecht and Stone 2015) use the network architecture most similar to ours. They use a fixed trace length parameter rather than our possession-based method. Hausknecht and Stone find that for partially observable processes, the LSTM mechanism outperforms a memory window. Our lesion study confirms this finding in an on policy prediction problem.

Player Ranking. (Albert et al. 2017) provide several up-to-date survey articles about player ranking. Previous work has taken one of two main approaches: 1) total action value counts (Decroos et al. 2017b), and 2) value-above-replacement. A value-above-replacement approach begins with a model that predicts a score of interest given player statistics. The performance metric for a specific player is then the difference between the score predicted given the player's statistics, and the score predicated given the statistics of an average or random player. For example, the Win-Above-Replacement (WAR) metrics compare the team's winning chances with the target player to the winning chances when the target player is replaced by an average player (Gerstenberg et al. 2014). To our knowledge, our Total Impact Metric is the only ranking metric that satisfies both the action count and replacement principles.

Reinforcement Learning and Player Ranking. Using the Q-function for player ranking is a recent development (Schulte et al. 2017; Cervone et al. 2014a; Routley and Schulte 2015). None of this previous work used a neural network. Schulte et al. discretized location and time coordinates and applied dynamic programming to learn a Q-function. Discretization leads to loss of information, undesirable spatio-temporal discontinuities in the Q-function, and generalizes poorly to unobserved parts of the state space. Another difference is that their action count model derived from the Q-function considers only the current state rather than the state transition as we do. Intuitively, this means that a player is given credit for doing the right thing at the right time, but not for being in the right place. In basketball, Cervone et al. defined a value-above-replacement metric with a expected point value model that is equivalent to a Q-function. They did not use reinforcement learning. Their approach assumes complete observability (of all players at all times), while our data provide partial observability only.

Data Description

The dataset was constructed by SportLogiq using computer vision techniques. The data provide information about game events and player actions for the entire 2015-2016 NHL season, which contains 3,382,129 events, covering 30 teams, 1140 games and 2,233 players. Table 1 shows an excerpt. The data track events around the puck, and records the identity and actions of the player in possession, with space and time stamps, and features of the game context. We used 13 of the recorded action types listed in the supplementary material. The table utilizes adjusted spatial coordinates where negative numbers refer to the defensive zone of the acting player, positive numbers to his offensive zone. Adjusted X-coordinates run from -100 to +100, Y-coordinates from 42.5 to -42.5, and the origin is at the ice center like Figure 2. We augment the data with derived features in Table 2 and list the complete feature set in Table 3.

Play Dynamics in the NHL

The **goal vector** g_t is a 1-of-3 indicator vector that specifies which team scores. For readability, we use $Home, Away, Neither$ to denote the team in a goal vector. For example, $g_{t,Home} = 1$ means that the home team scores at time t. A **feature vector** $\mathbf{x}_t$ for discrete time step t specifies a value for each of the 10 features listed in Table 3, and for each goal indicator. The **action** a_t is one of the 13 types, together with a mark that specifies the team executing the action, e.g. $Shot(Home)$. A **sequence** s_t is a list $(\mathbf{x}_0, a_0, \ldots, \mathbf{x}_t)$ of observations and actions.

We can apply reinforcement learning methods for Markov processes, simply by using the complete sequence s_t as the state representation at time step t (Mnih et al. 2015). It is well-known that in the on-policy setting, policy and environment can be combined into a single Markov reward process model (Sutton and Barto 1998) that maps state-action pairs to a distribution over subsequent state-action-reward triples: $\sigma(s_{t+1}, a_{t+1}, g_{t+1}|s_t, a_t)$.

Goals and the Q-function in NHL Play. We divide NHL games into goal-scoring episodes, so that each episode 1) begins at the beginning of the game, or immediately after a goal, and 2) terminates with a goal or the end of the game. We define a separate Q function for the three outcomes:

$$Q_{team}(s, a) = P_{\sigma}(goal_{team} = \mathit{1}|s_t = s, a_t = a)$$

represents the conditional probability that the home resp. away team scores the goal at the end of the current episode (denoted $goal_{Home} = \mathit{1}$ resp. $goal_{Away} = \mathit{1}$), or neither team does (denoted $goal_{Neither} = \mathit{1}$).

Table 1: Dataset Example

GID	PID	GT	TID	X	Y	MP	GD	Action	OC	P
1365	126	14.3	6	-11.0	25.5	Even	0	Lpr	S	A
1365	126	17.5	6	-23.5	-36.5	Even	0	Carry	S	A
1365	270	17.8	23	14.5	35.5	Even	0	Block	S	A
1365	126	17.8	6	-18.5	-37.0	Even	0	Pass	F	A
1365	609	19.3	23	-28.0	25.5	Even	0	Lpr	S	H
1365	609	19.3	23	-28.0	25.5	Even	0	Pass	S	H

Table 2: Derived Features

Velocity	TR	D	Angle	H/A	PN
(-23.4, 1.5)	3585.7	3.4	0.250	A	4
(-4.0, -3.5)	3582.5	3.1	0.314	A	4
(-27.0, -3.0)	3582.2	0.3	0.445	H	4
(0, 0)	3582.2	0.0	0.331	A	4
(-30.3, -7.5)	3580.6	1.5	0.214	H	5
(0,0)	3580.6	0.0	0.214	H	5

GID=GameId, PID=playerId, GT=GameTime, TID=TeamId, MP=Manpower, GD=Goal Difference, OC = Outcome, S=Succeed, F=Fail, P = Team Possess puck, H=Home, A=Away, H/A=Team who performs action, TR = Time Remain, PN = Play Number, D = Duration

Table 3: Complete Feature List

Name	**Type**	**Range**
X Coordinate of Puck	Continuous	[-100, 100]
Y Coordinate of Puck	Continuous	[-42.5, 42.5]
Velocity of Puck	Continuous	(-inf, +inf)
Game Time Remain	Continuous	[-300, 3600]
Score Differential	Discrete	(-inf, +inf)
Manpower Situation	Discrete	{EV, SH, PP}
Event Duration	Continuous	[0, +inf)
Action Outcome	Discrete	{successful, failure}
Angle between puck and goal	Continuous	$[-3.14, 3.14]$
Home or Away Team	Discrete	{Home, Away}

One of the advantages of reinforcement learning is that it can be applied with different rewards of interest, such as match win (Pettigrew 2015; Kaplan, Mongeon, and Ryan 2014; Routley 2015) and penalties (Routley and Schulte 2015). For player ranking, the next goal reward has two advantages over match win. 1) The next goal reward is closer to what a coach expects from a player. For example if a team is ahead by two goals with one minute left in the match, a player's actions have negligible effect. Nonetheless professionals should keep playing as well as they can and maximize the scoring chances for their own team. 2) A single action has a small effect on the final outcome, leading to numerically close performance metrics for the players.

Illustration of Network Predictions

Before we describe our training method, we illustrate the representational power of the trained network. The network utilizes an LSTM architecture that takes a current sequence s_t, an action a_t as input, and is trained to predicted the expected returns. There are three output nodes, representing the estimates $\hat{Q}_{Home}(s, a)$, $\hat{Q}_{Away}(s, a)$ and $\hat{Q}_{Neither}(s, a)$. Output values are normalized to be probabilities. The Q-functions for each team share network weights and therefore similar features extracted in hidden layers.

To provide a qualitative sense of the predictions, we give examples that illustrate both the spatial and the temporal dimensions of our learned neural net.

Temporal Projection. Figure 1 shows a value ticker (Decroos et al. 2017a; Cervone et al. 2014b) that represents the evolution of the action-value function throughout a randomly selected match, Penguins vs Canadiens, Oct 13, 2015. The figure plots values of the three output nodes. Sports analysts and commentators use ticker plots to highlight critical match events. We mark significant changes in the scoring probabilities (action-value Q) and explain their corresponding events in Table 4.

Table 4: Ticker Event Highlights

Label	**What happened**
1	Canadiens shot near Penguins' goal
2	Canadiens passed around Penguins' goal, shot
3	Penguins shot (failed to score), recovered the puck, passed around the goal, Canadiens blocked
4	Penguins received penalty Canadiens scored on the powerplay

Spatial Projection. Because the neural network generalizes from observed sequences and actions to sequences and actions that have not occurred in the observed NHL season, we can plot an action-value function for the entire rink. Figure 2 shows the learned smooth value surface $\hat{Q}_{Home}(s, shot(team))$ for home team shots. We provide one heat map with no previous history, and another for a representative play sequence, reception-pass-reception; these are the most frequent home actions preceding a home shot. Other features are filled in using the data mean; for complete details please see the supplementary material.

The DP-LSTM Design

We describe our network architecture and training method. **Network Architecture.** A 5-layer network with 3 hidden layers. Each hidden layer contains 1000 nodes, which utilize a relu activation function. The first hidden layer is the LSTM layer, the remaining layers are fully connected.

Possession-Based Dynamic Trace Length. The key trace length parameter controls how far back in time the LSTM propagates the error signal from the current time at the input history. We break a goal-scoring episode into further sub-episodes called plays. A **play** can end with the loss of possession to the other team, but also with a play stoppage such as face-offs, penalties, and period breaks. The trace length is then set to go back to the beginning

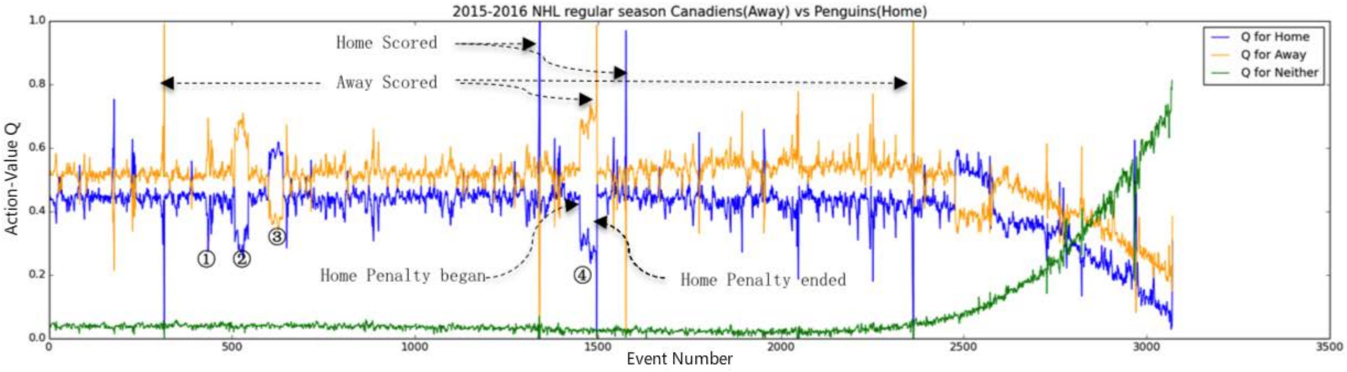

Figure 1: Temporal Projection of method. Changes in the scoring probabilities result from major events as shown. The network also captures smaller changes associated with every action. We find 1) High scoring opportunity of one team will inhibit that of its opponent and 2)The probability that neither team scores rises significantly at the end of the match.

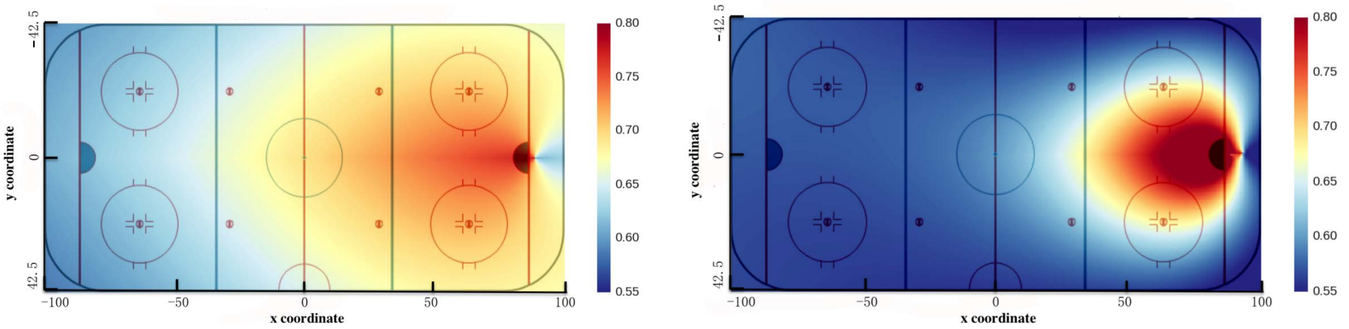

Figure 2: Spatial Projection of method. **Left:** The shot-value function with no history. **Right:** The shot-value function after a history of three actions by the home team. It can be observed that 1) The chance that the home team scores after a shot is shown to depend on the angle and distance to the goal. 2) Action-value function is generalized to regions where shots are rarely observed, for instance close to the middle of the rink. 3) With history, DP-LSTM become more confident to predict larger scoring chance around opponent's goal and less scoring chance away from it. 4)The LSTM captures the increase in goal scoring probability that results from the home team's momentum.

of the current play (possession episode). Using possession changes to define episodes is common in several continuous-flow sports, especially basketball (Cervone et al. 2014a; Omidiran 2011). Figure 3 illustrates the concept. Table 2 shows two plays and one possession change, from Play 4 to Play 5 at game time 17.8 seconds. The motivation is that because of the change in puck possession, we expect a high correlation for sequence values within a play and a low correlation for sequence values from different plays. The average play is fairly short with 4.65 actions, which facilitates fast and accurate LSTM backpropgation training. There is however considerable variance with the maximum length at 43, which is why we do not simply use a short fixed-length trace. To prevent LSTM trace from becoming too long, we set the maximum trace length to 10 (as in (Hausknecht and Stone 2015)).

Table 5: Cost of Neural Network, MCAVE = Monte Carlo Action Value Estimation, TD-Sarsa = Temporal Differen Sarsa

Training Methods	**Error Signal**
MCAVE	$(\sum_{n=t} g_n - \hat{Q}(s_t, a_t))^2$
TD-Sarsa	$(g_{t+1} + \hat{Q}(s_{t+1}, a_{t+1}) - \hat{Q}(s_t, a_t))^2$

Weight Training. We apply the Sarsa algorithm (Sutton and Barto 1998)[Ch.6.4], a TD learning method for estimating Q values. Table 5 shows the TD(0)-Sarsa error function for weight training. For *data preprocessing,* we standardize input features to have zero mean and unit variance. Weights are optimized by backpropagation using minibatch gradient descent with batch size 32 (determinally experimentally). Backpropagation requires an initial set of weights. For *weight initialization*, we introduce a novel simple average-input average-output pre-training method: The net's starting point maps the average input sequence to the average scoring chance vector. The input vector is the average of all concatenated feature vectors. The output vector is the average of all

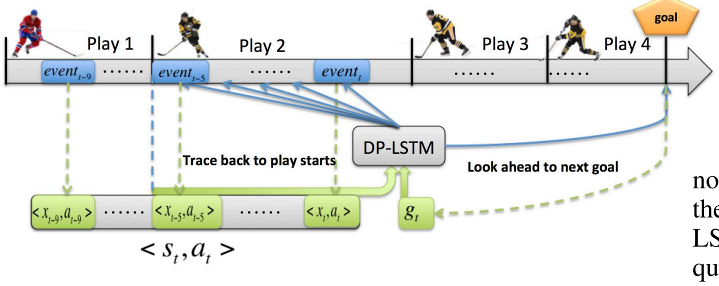

Figure 3: Within our DP-LSTM design, temporal-difference learning looks ahead to the next goal, and the LSTM memory traces back to the beginning of the play (the last possession change).

scoring chance vectors $goal^{obs}_{team}(s)$ over the length 10 sequences. These average scoring chances are 0.462 for home and 0.424 for away, and 0.114 for neither team. The neural net is then trained to predict the mean output vector for the mean input vector with error below 10^{-4}.

Empirical Evaluation

We show that the DP-LSTM design converges quickly and smoothly. Also, the trained LSTM is well-calibrated, in the sense that for the sequences in actual NHL plays, the probabilities assigned by DP-LSTM match the observed event frequencies well. We compare our method to a set of lesion methods that remove a component of the full model.

Calibration and Data Fit

To define empirical event frequencies, we discretize the continuous sequence space into discrete bins. A sequence is assigned to a bin according to the values of three discrete *context features* observed in the last sequence stage: period (1, 2, 3), manpower differential(shorthanded, even strength, power play) and score differential (-3, -2, -1, 0, 1, 2, 3). So the total number of sequence bins is $3 \times 3 \times 7 = 63$. This partition has two advantages. 1) The context features are well-studied and important for hockey experts (Thomas et al. 2013; Routley and Schulte 2015; Pettigrew 2015; Kaplan, Mongeon, and Ryan 2014), so the model predictions can be checked against domain knowledge. For example, manpower advantage leads to a higher scoring chance (Thomas et al. 2013; Routley and Schulte 2015). Less obvious is the phenomenon of home team advantage: Comparing two match contexts with the home and away team roles exchanged, the relative advantage of the home team is greater than that of the away team (Swartz and Arce 2014; Routley and Schulte 2015). 2) The partition covers a wide range of match contexts, and each bin aggregates a large set of play histories. If our model exhibits a systematic bias, the aggregation should amplify it to become detectable.

The empirical and model goal distributions to be compared are defined as follows. For each observed sequence s, set $goal^{obs}_{team}(s) = 1$ if the (unique) observed episode containing s ends with a goal by team $team = Home, Away$ or neither ($team = Neither$). Then

$$Q^{obs}_{team}(A) = \frac{1}{|A|} \sum_{s \in A} goal^{obs}_{team}(s)$$

where A represents one of 63 sequence bins and $|A|$ denotes the number of sequences observed in A. To compute the estimated goal scoring distribution, we apply the DP-LSTM to compute action-value Q for each observed sequence and average the resulting estimates:

$$\hat{Q}_{team}(A) = \frac{1}{|A|} \sum_{s \in A} \hat{Q}_{team}(s,a).$$

We compare the distributions using the Kullback-Leibler divergence (KLD), defined as $Q_1||Q_2 \equiv \sum_x Q_1(x) \ln \frac{Q_1(x)}{Q_2(x)}$. Table 6 shows the probability differences for some bins with match contexts. We selected the 11 most frequent contexts and 6 most frequent contexts with some man power differential to evaluate home team advantage. The data fit is excellent, below KLD of 0.06 except for the relatively rare situation where the away team has a manpower advantage in a tied third period.

The estimated value function matches several well-known phenomena. 1) Manpower advantage has the biggest impact, necessary to get a scoring chance above 60%. 2) Scoring rates decrease in later periods. 3) A clear home team advantage. For example, a manpower advantage of 1 in the first period translates into a high scoring chance estimate of 69% for the home team, but only 63% for the away team.

Table 6: Observed vs. Predicted Scoring Chances

GD	MD	P	#N	H	A	KLD
0	0	1	294602	0.53	0.47	0.0269
0	0	2	135138	0.52	0.48	0.0010
0	0	3	103590	0.37	0.34	0.0147
-1	0	2	97978	0.52	0.48	0.0027
1	0	2	97054	0.53	0.47	0.0073
-1	0	3	93807	0.36	0.34	0.0016
1	0	1	79980	0.54	0.46	0.0105
1	0	3	79941	0.37	0.35	0.0127
-1	0	1	77195	0.53	0.47	0.0081
-2	0	3	59797	0.32	0.31	0.0073
2	0	3	54605	0.34	0.32	0.0215
0	1	1	39692	0.69	0.31	0.0202
0	-1	1	11384	0.37	0.63	0.0144
0	1	2	25724	0.67	0.33	0.0100
0	-1	2	7337	0.37	0.63	0.0057
0	1	3	13116	0.54	0.22	0.0558
0	-1	3	3430	0.25	0.54	0.0912

GD=Goal Differential, MD=Manpower Differential, P=Period, #Sequences, H=Predicted Home Value, A=Predicted Away Value

Lesion Study

We compare the full DP-LSTM with alternative lesion designs that replace or omit components. **FT-LSTM** uses a fixed length trace parameter $tl = 10$. Besides, instead of inputting an entire sequence up to the current time t, an architecture with a *fixed window size* concatenates the last k

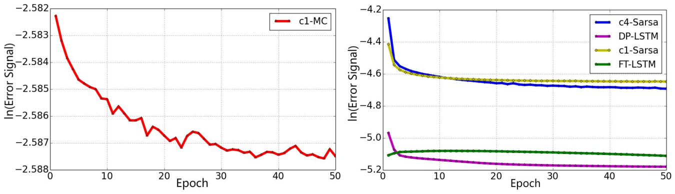

Figure 4: Average error signal within 50 iterations for **left**:c1-MC **right**:c1-Sarsa, c4-Sarsa, FT-LSTM and DP-LSTM.

feature-action vectors as input. With $k = 1$, only the current action and features are presented to the neural net. The fixed window-methods can be implemented using neural nets with fixed input sizes, where we replace the LSTM layer with a fully connected FCC layer. We refer to learning with the window size k as ck-x. For example, **c4-Sarsa** applies TD-learning with a history window of size 4. This design is similar to that introduced in (Mnih et al. 2015). We also evaluate **c1-Sarsa** to assess the importance of including play history in the input.

For the Monte Carlo error function (see Table 5), we use a fixed window size with $k = 1$ and the observed goal vector as the target value vector without looking ahead. We refer to this method as **c1-MC**. Table 7 summarizes the evaluated designs.

Table 7: Neural Network Architecture

Name	Training Methods	Function Approximation
c1-MC	MCAVE	FCNN
c1-Sarsa	TD-Sarsa	FCNN
c4-Sarsa	TD-Sarsa	c4 input FCNN
FT-LSTM	TD-Sarsa	Fixed Trace Length LSTM +FCNN
DP-LSTM	TD-Sarsa	Dynamic Play Trace LSTM + FCNN

Results. We report results about the convergence of weight learning, and about the quality of the learned weights, measured by weighted KLD. The measurements report were obtained with the following system settings. We ran Tensorflow 1.1 on a TITAN X GPU with 12GB Main Memory. We pass over the entire training dataset 50 times (epochs). We applied the ADAM optimizer for adapting learning rates (Kingma and Ba 2014), with initial learning rate 10^{-5} (determined experimentally by grid search) . For our full DP-LSTM model, training took 5 days computation time and over 5 million minibatch gradient descent steps.

Training Error Convergence. The error of neural networks is defined in table 5. Figure 4 plots the average ln training error in each epoch. We observe the following. 1) The two LSTM errors are lower that for the three fixed-length designs. DP-LSTM reaches the lowest error overall within 10 epochs. 2) The average error for c1-Sarsa, c4-Sarsa and DP-LSTM drops substantially in the first 20 epochs. 3) The error signal for FT-LSTM increases at first, because the network is exposed to more irrelevant actions. This shows the value of building in the concept of possessions (plays) via the trace length.

Data Fit. To aggregate the context results into a single metric, we average the KLD of bins by their size using the **weighted KLD** metric:

$$\overline{KLD}(Q_{obs}, \hat{Q}) = \sum_{A} \frac{|A|}{N} Q_{obs} || \hat{Q}$$

where N is the total number of observed play sequences. The bin sizes are highly variable, as the *#N* column of Table 6 illustrates. The observations in larger bins more are likely to represent the dynamics of the underlying game process. The weighted KLD for DP-LSTM is 1.39×10^{-2}. Table 8 show the resulting weighted KLD. DP-LSTM shows the closest fit to the empirical frequencies, with a similar value for c4-Sarsa, which shows that a window size of 4 captures much but not all relevant play history. FT-LSTM has a worse score, which illustrates the drawback of fixing trace length. The memoryless designs c1-MC has the worst fit, showing the value of bootstrapping and play history.

Table 8: Weighted KLD.

c1-MC	c1-Sarsa	c4-Sarsa	FT-LST	**DP-LSTM**
7.93×10^{-2}	1.94×10^{-2}	1.93×10^{-2}	2.18×10^{-2}	$\mathbf{1.39 \times 10^{-2}}$

Player Ranking

We show how to quantify the impact of a player's actions with the learned action values and then rank players by the resulting impact performance metric.

Total Action Values Based on the Q-function We measure the quality of an action by how much it changes the expected return of a player's team. In terms of the Q-function, this is the *change in Q-value* due to a player's action. This quantity is defined as the action's **impact**. The impact can be visualized as the difference between successive points in the Q-value ticker (Figure 1). For our specific choice of Next Goal as the reward function, we refer to **goal impact**. Goal impact captures a player's offensive contributions—increasing the scoring chances of his team—as well as his defensive contributions—decreasing the scoring chances of the opposing team.The total impact of a player's actions is his **Goal Impact Metric** (GIM). The formal equations are:

$$impact^{team}(s_t, a_t) = Q^{team}(s_t, a_t) - Q^{team}(s_{t-1}, a_{t-1})$$
$$GIM^i(D) = \textstyle\sum_{s,a} n^i_D(s, a) \times impact^{team_i}(s, a)$$

where D indicates our dataset, $team_i$ denotes the team of player i, and $n^i_D(s, a)$ is the number of times that player i was observed to perform action a at s.

Q-Value Above Replacement We introduce the **Q-value-above-replacement** (QAR) metric that replaces at a game history the acting player with a random player and computes the resulting difference in expected returns. The **transition probabilities** for a player i include the probability that player i acts in the next state:

$$\sigma_i(s_{t+1}, a_{t+1}|s_t, a_t) \equiv Pr(s_{t+1}, a_{t+1}, player_{t+1} = i|s_t, a_t),$$

with maximum likelihood estimate $\hat{\sigma}_i(s_{t+1}, a_{t+1}|s_t, a_t) \equiv \frac{n^i_D(s_{t+1}, a_{t+1}, s_t, a_t)}{n^i_D(s_t, a_t)}$. The pooled process is a mixture of the individual processes: $\sigma = \sum_i \sigma_i$ and can be interpreted as specifying the state transition probabilities for a random player. Given that player i acts next, we can compute the expected Q-value at time t, and compare it to the expected Q-value of a random player to define the QAR metric:

$$Q^i(s_t, a_t) \equiv \sum_{s,a} Q^{team_i}(s, a) \times \sigma_i(s_{t+1} = s, a_{t+1} = a|s_t, a_t)$$

$$QAR^i(D) \equiv \sum_{s,a} n^i_D(s, a) \times [Q^i(s, a) - Q^{team_i}(s, a)]$$

We prove that total goal impact and Q-value-above-replacement are equivalent, which provides a theoretical foundation for our metric. The proof is in the supplement.

Proposition 1 *For the maximum likelihood estimate $\hat{\sigma}_i$ we have $QAR^i(D) = GIM^i(D)$.*

Empirical Results

We discuss individual players that highlight special features of our GIM metric. For quantitative evaluation, we show high correlation of GIM with previously established metrics. High temporal auto-correlation establishes that the metric is temporally consistent.

Case Studies. The top-20 highest impact with player are listed in Table 9. All these players are well-known NHL stars. Taylor Hall tops the ranking although he did not score the most goals. This shows how our ranking, while correlated with goals, also *reflects the value of other actions by the player.* For instance, our ranking reflects that the total number of Hall's passes is exceptionally high at 320.

Table 9: 2015-2016 Top-20 Player Impact Scores

Name	Impact	Assists	Goals	Points	+/-	Salary
Taylor Hall	96.40	39	26	65	-4	$6,000,000
Joe Pavelski	94.56	40	38	78	25	$6,000,000
Johnny Gaudreau	94.51	48	30	78	4	$925,000
Anze Kopitar	94.10	49	25	74	34	$7,700,000
Erik Karlsson	92.41	66	16	82	-2	$7,000,000
Patrice Bergeron	92.06	36	32	68	12	$8,750,000
Mark Scheifele	90.67	32	29	61	16	$832,500
Sidney Crosby	90.21	49	36	85	19	$12,000,000
Claude Giroux	89.64	45	22	67	-8	$9,000,000
Dustin Byfuglien	89.46	34	19	53	4	$6,000,000
Jamie Benn	88.38	48	41	89	7	$5,750,000
Patrick Kane	87.81	60	46	106	17	$13,800,000
Mark Stone	86.42	38	23	61	-4	$2,250,000
Blake Wheeler	85.83	52	26	78	8	$5,800,000
Tyler Toffoli	83.25	27	31	58	35	$2,600,000
Charlie Coyle	81.50	21	21	42	1	$1,900,000
Tyson Barrie	81.46	36	13	49	-16	$3,200,000
Jonathan Toews	80.92	30	28	58	16	$13,800,000
Sean Monahan	80.92	36	27	63	-6	$925,000
Vladimir Tarasenko	80.68	34	40	74	7	$8,000,000

Our metric can be used to *identify undervalued players.* For instance, Johnny Gaudreau and Mark Scheifele drew salaries below what their GIM rank would suggest. Later they received a $\$5M+$ contract for the 2016-17 season.

Correlations with Standard Metrics. We compare goal impact with four well-known player statistics: assists, goals, points (goals + assists), +/-. Correlations between GIM and metrics like Assists, Goals, Points and Impact are high (0.87, 0.88 and 0.93 respectively). The high correlations with known player metrics provides support for the goal impact metric. The GIM metric is quite *different from the +/- score* ($\rho = 0.15$) because +/- reflects the performance of all players on the ice rather than a player's individual achievements. The strong correlation with points suggests a potential application for fantasy play. In a fantasy league, fans select players for a season and their "team" is scored on its total number of points in the season. Millions of fans participate, and the NHL provides extensive advice and player rankings. For instance, Joe Pavelski and Erik Karlsson place at ranks 3 and 6 respectively in the NHL fantasy rankings.

Auto-Correlation. A desirable feature of a player ranking score is temporal consistency (Pettigrew 2015), because it means that future performance can be predicted from past performance. GIM is a stable metric; after half the season the correlation between the observed goal impact so far and the final goal impact is already above 0.9. The supplementary material plots the temporal auto-correlation of GIM.

Conclusion and Future Work

We applied Deep Reinforcement Learning to model complex spatio-temporal NHL dynamics. The trained neural network provides a rich source of knowledge about how a team's chance of scoring the next goal (Q-value) depends on the match context, the rink location, and the game time. From this knowledge we derived a performance metric that measures how a player's actions affect his team's scoring chances.

Our DP-LSTM design incorporates the insight that team sports have a turn-taking aspect where one team is on the offensive and the other defends. So we restrict history traces to remain within the same offensive play. Evaluated on a massive complex dataset comprising over 3M NHL events, DP-LSTM learning shows excellent convergence and data fit, better than four comparison lesion methods. The proposed player performance metric summarizes rich performance data in an informative metric that is temporally consistent and facilitates identifying undervalued players.

Avenues for future work include the following. 1) Sports is a natural application domain for hierarchical Reinforcement Learning (Dietterich 2000), which decomposes winning into smaller tasks like defensing opponent or reducing penalties. 2) Explainability is important for decision support (e.g., advising coaches). A common method for making neural net predictions transparent is to translate them into a tree model (Johansson, Sönströd, and König 2014). This raises the challenge of extracting tree models from a sports LSTM.

The techniques we have developed can be transferred to other physical games. In sports games we have large datasets, modelling challenges for complex event data, and intense interest in analytical insights from both participants and the public. Sports analytics has the potential to become a major application area for deep reinforcement learning.

References

Albert, J.; Glickman, M. E.; Swartz, T. B.; and Koning, R. H. 2017. *Handbook of Statistical Methods and Analyses in*

Sports. CRC Press.

Cervone, D.; DAmour, A.; Bornn, L.; and Goldsberry, K. 2014a. Pointwise: Predicting points and valuing decisions in real time with NBA optical tracking data. In *8th Annual MIT Sloan Sports Analytics Conference, February*, volume 28.

Cervone, D.; DAmour, A.; Bornn, L.; and Goldsberry, K. 2014b. Pointwise: Predicting points and valuing decisions in real time with nba optical tracking data. In *8th Annual MIT sloan sports analytics conference, February*, volume 28.

Decroos, T.; Dzyuba, V.; Haaren, J. V.; and Davis, J. 2017a. Predicting soccer highlights from spatio-temporal match event streams. In *AAAI 2017*, 1302–1308.

Decroos, T.; Van Haaren, J.; Dzyuba, V.; and Davis, J. 2017b. Starss: A spatio-temporal action rating system for soccer. In *Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2017 workshop*.

Dietterich, T. G. 2000. Hierarchical reinforcement learning with the maxq value function decomposition. *J. Artif. Intell. Res.(JAIR)* 13:227–303.

Gerstenberg, T.; Ullman, T.; Kleiman-Weiner, M.; Lagnado, D.; and Tenenbaum, J. 2014. Wins above replacement: Responsibility attributions as counterfactual replacements. In *Proceedings of the Cognitive Science Society*, volume 36.

Hausknecht, M., and Stone, P. 2015. Deep recurrent q-learning for partially observable mdps. *CoRR, abs/1507.06527*.

Johansson, U.; Sönströd, C.; and König, R. 2014. Accurate and interpretable regression trees using oracle coaching. In *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on*, 194–201. IEEE.

Kaplan, E. H.; Mongeon, K.; and Ryan, J. T. 2014. A markov model for hockey: Manpower differential and win probability added. *INFOR: Information Systems and Operational Research* 52(2):39–50.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Macdonald, B. 2011. A regression-based adjusted plus-minus statistic for nhl players. *Journal of Quantitative Analysis in Sports* 7(3):29.

McHale, I. G.; Scarf, P. A.; and Folker, D. E. 2012. On the development of a soccer player performance rating system for the english premier league. *Interfaces* 42(4):339–351.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

Omidiran, D. 2011. A new look at adjusted plus/minus for basketball analysis. In *MIT Sloan Sports Analytics Conference [online]*.

Pettigrew, S. 2015. Assessing the offensive productivity of nhl players using in-game win probabilities. In *9th Annual MIT Sloan Sports Analytics Conference*.

Routley, K., and Schulte, O. 2015. A markov game model for valuing player actions in ice hockey. In *Uncertainty in Artificial Intelligence (UAI)*, 782–791.

Routley, K. 2015. A markov game model for valuing player actions in ice hockey. Master's thesis, Simon Fraser University.

Schuckers, M., and Curro, J. 2013. Total hockey rating (thor): A comprehensive statistical rating of national hockey league forwards and defensemen based upon all on-ice events. In *7th Annual MIT Sloan Sports Analytics Conference*.

Schulte, O.; Khademi, M.; Gholami, S.; Zhao, Z.; Javan, M.; and Desaulniers, P. 2017. A markov game model for valuing actions, locations, and team performance in ice hockey. *Data Mining and Knowledge Discovery* 1–23.

Spagnola, N. 2013. The complete plus-minus: A case study of the columbus blue jackets. Master's thesis, University of South Carolina.

Sutton, R. S., and Barto, A. G. 1998. *Introduction to reinforcement learning*, volume 135. MIT Press Cambridge.

Swartz, T. B., and Arce, A. 2014. New insights involving the home team advantage. *International Journal of Sports Science & Coaching* 9(4):681–692.

Thomas, A.; Ventura, S.; Jensen, S.; and Ma, S. 2013. Competing process hazard function models for player ratings in ice hockey. *The Annals of Applied Statistics* 7(3):1497–1524.