Deep Reinforcement Learning in Ice Hockey
for Context-Aware Player Evaluation
Guiliang Liu and Oliver Schulte
Simon Fraser University, Burnaby, Canada
gla68@sfu.ca, oschulte@cs.sfu.ca

Abstract
A variety of machine learning models have been proposed to assess the performance of players in professional sports. However, they have only a limited ability to model how player performance depends on the game context. This paper proposes a new approach to capturing game context: we apply Deep Reinforcement Learning (DRL) to learn an action-value Q function from 3M play-by-play events in the National Hockey League (NHL). The neural network representation integrates both continuous context signals and game history, using a possession-based LSTM. The learned Q-function is used to value players’ actions under different game contexts. To assess a player’s overall performance, we introduce a novel Game Impact Metric (GIM) that aggregates the values of the player’s actions. Empirical Evaluation shows GIM is consistent throughout a play season, and correlates highly with standard success measures and future salary.

1 Introduction: Valuing Actions and Players
With the advancement of high frequency optical tracking and object detection systems, more and larger event stream datasets for sports matches have become available. There is increasing opportunity for large-scale machine learning to model complex sports dynamics. Player evaluation is a major task for sports modeling that draws attention from both fans and team managers, who want to know which players to draft, sign or trade. Many models have been proposed [Buttrey et al., 2011; Macdonald, 2011; Decroos et al., 2018; Kaplan et al., 2014]. The most common approach has been to quantify the value of a player’s action, and to evaluate players by the total value of the actions they took [Schuckers and Curro, 2013; McHale et al., 2012].

However, traditional sports models assess only the actions that have immediate impact on goals (e.g. shots), but not the actions that lead up to them (e.g. pass, reception). And action values are assigned taking into account only a limited context of the action. But in realistic professional sports, the relevant context is very complex, including game time, position of players, score and manpower differential, etc.

Recently, Markov models have been used to address these limitations. [Routley and Schulte, 2015] used states of a Markov Game Model to capture game context and compute a Q function, representing the chance that a team scores the next goal, for all actions. [Cervone et al., 2014] applied a competing risk framework with Markov chain to model game context, and developed EPV, a point-wise conditional value similar to a Q function, for each action. The Q-function concept offers two key advantages for assigning values to actions [Schulte et al., 2017a; Decroos et al., 2018]: 1) All actions are scored on the same scale by looking ahead to expected outcomes. 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.

The states in the previous Markov models represent only a partial game context in the real sports match, but nonetheless the models assume full observability. Also, they pre-discretized input features, which leads to loss of information. In this work, we utilize a deep reinforcement learning (DRL) model to learn an action-value Q function for capturing the current match context. The neural network representation can easily incorporate continuous quantities like rink location and game time. To handle partial observability, we introduce a possession-based Long Short Term Memory (LSTM) architecture that takes into account the current play history. Unlike most previous work on active reinforcement learning (RL), which aims to compute optimal strategies for complex continuous-flow games [Hausknecht and Stone, 2015;
we solve a prediction (not control) problem in the passive learning (on policy) setting [Sutton and Barto, 1998]. We use RL as a behavioral analytics tool for real human agents, not to control artificial agents.

Given a Q-function, the impact of an action is the change in Q-value due to the action. Our novel Goal Impact Metric (GIM) aggregates the impact of all actions of a player. To our knowledge, this is the first player evaluation metric based on DRL. The GIM metric measures both players’ offensive and defensive contribution to goal scoring. For player evaluation, similar to clustering, ground truth is not available. A common methodology [Routley and Schulte, 2015; Pettigrew, 2015] is to assess the predictive value of a player evaluation metric for standard measures of success. Empirical comparison between 7 player evaluation metrics finds that 1) given a complete season, GIM correlates the most with 12 standard success measures and is the most temporally consistent metric, 2) given partial game information, GIM generalizes best to future salary and season total success.

2 Related Work

We discuss the previous work most related to our approach.

Deep Reinforcement Learning. Previous DRL work has focused on control in continuous-flow games, not prediction [Mnih et al., 2015]. Among these papers, [Hausknecht and Stone, 2015] use a very similar network architecture to ours, but with a fixed trace length parameter rather than our possession-based method. Hausknecht and Stone find that for partially observable control problems, the LSTM mechanism outperforms a memory window. Our study confirms this finding in an on policy prediction problem.

Player Evaluation. Albert et al. 2017 provide several up-to-date survey articles about evaluating players. A fundamental difficulty for action value counts in continuous-flow games is that they traditionally have been restricted to goals and actions immediately related to goals (e.g. shots). The Q-function solves this problem by using lookahead to assign values to all actions.

Player Evaluation with Reinforcement Learning. Using the Q-function to evaluate players is a recent development [Schulte et al., 2017a; Cervone et al., 2014; Routley and Schulte, 2015]. 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. For basketball, Cervone et al. defined a player performance metric based on an expected point value model that is equivalent to a Q-function. Their approach assumes complete observability (of all players at all times), while our data provide partial observability only.

3 Task Formulation and Approach

Player evaluation (the “Moneyball” problem) is one of the most studied tasks in sports analytics. Players are rated by their observed performance over a set of games. Our approach to evaluating players is illustrated in Figure 2. Given dynamic game tracking data, we apply Reinforcement Learning to estimate the action value function $Q(s, a)$, which assigns a value to action $a$ given game state $s$. We define a new player evaluation metric called Goal Impact Metric (GIM) to value each player, based on the aggregated impact of their actions, which is defined in Section 6 below. Player evaluation is a descriptive task rather than a predictive generalization problem. As game event data does not provide a ground truth rating of player performance, our experiments assess player evaluation as an unsupervised problem in Section 7.

4 Play Dynamic in NHL

We utilize a dataset constructed by SPORTLOGiQ using computer vision techniques. The data provide information about game events and player actions for the entire 2015-2016 NHL (largest professional ice hockey league) 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 record the identity and actions of the player in possession, with space and time stamps, and features of the game context. 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 as in Figure 1. We augment the data with derived features in Table 2 and list the complete feature set in Table 3.

We apply the Markov Game framework [Littman, 1994] to learn an action value function for NHL play. Our notation for RL concepts follows [Mnih et al., 2015]. There are two agents Home resp. Away representing the home resp. away team. The reward, represented by goal vector $g$, is a 1-of-3 indicator vector that specifies which team scores (Home, Away, Neither). An action $a_t$ is one of 13 types, including shot, block, assist, etc., together with a mark that specifies the team executing the action, e.g. Shot(Home). An observation is a feature vector $x_t$ for discrete time step $t$ that specifies a value for the 10 features listed in Table 3. We use the complete sequence $s_t = (x_t, a_{t-1}, x_{t-1}, \ldots, x_0)$ as the state representation at time step $t$ [Mnih et al., 2015], which satisfies the Markov property.

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. A Q function represents the conditional probability of the event that the home resp. away team scores the goal at the end of the current episode (denoted $\text{goal}_{\text{Home}} = 1$ resp. $\text{goal}_{\text{Away}} = 1$), or neither team does (denoted $\text{goal}_{\text{Neither}} = 1$):
where team is a placeholder for one of Home, Away, Neither. This Q-function represents the probability that a team scores the next goal, given current play dynamics in the NHL (cf. Schulte et al.; Routley and Schulte). Different Q-functions for different expected outcomes have been used to capture different aspects of NHL play dynamics, such as match win [Pettigrew, 2015; Kaplan et al., 2014; Routley and Schulte, 2015] and penalties [Routley and Schulte, 2015]. For player evaluation, the next-goal Q function has three advantages. 1) The next-goal reward captures 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 on final match outcome. Nonetheless professionals should keep playing as well as they can and maximize the scoring chances for their own team. 2) The Q-values are easy to interpret, since they model the probability of an event that is a relatively short time away (compared to final match outcome). 3) Increasing the probability that a player’s team scores the next goal captures both offensive and defensive value. For example, a defensive action like blocking a shot decreases the probability that the other team will score the next goal, thereby increasing the probability that the player’s own team will score the next goal.

5 Learning Q values with DP-LSTM Sarsa

We take a function approximation approach and learn a neural network that represents the Q-function \( Q_{team}(s, a) \).

Table 1: Dataset Example

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

Table 2: Derived Features

![Figure 3: Our design is 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. 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).](image)

5.1 Network Architecture

Figure 3 shows our model structure. Three output nodes represent the estimates \( Q_{Home}(s, a) \), \( Q_{Away}(s, a) \) and \( Q_{Neither}(s, a) \). Output values are normalized to probabilities. The \( Q \)-functions for each team share weights. The network architecture is a Dynamic LSTM that takes as inputs a current sequence \( s_t \), an action \( a_t \) and a dynamic trace length \( tl_t \).¹

5.2 Weight Training

We apply an on-policy Temporal Difference (TD) prediction method Sarsa [Sutton and Barto, 1998, Ch.6.4], to estimate \( Q_{team}(s, a) \) for the NHL play dynamics observed in our dataset. Weights \( \theta \) are optimized by minibatch gradient descent via backpropagation. We used batch size 32 (determined experimentally). The Sarsa gradient descent update at time step \( t \) is based on a squared-error loss function:

¹We experimented with a single-hidden layer, but weight training failed to converge.
6 Player Evaluation

In this section, we define our novel Goal Impact Metric and give an example player ranking.

6.1 Player Evaluation Metric

Our Q-function concept provides a novel AI-based definition for assigning a value to an action. Like [Schulte et al., 2017b], we measure the quality of an action by how much it changes the expected return of a player’s team. Whereas the scoring chance at a time measures the value of a state, and therefore depends on the previous efforts of the entire team, the change in value measures directly the impact of an action by a specific player. 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 4). For our specific choice of Next Goal as the reward function, we refer to goal impact. The total impact of a player’s actions is his Goal Impact Metric (GIM). The formal equations are:

$$\text{impact}^{\text{team}}(s_t, a_t) = Q^{\text{team}}(s_t, a_t) - Q^{\text{team}}(s_{t-1}, a_{t-1})$$

$$\text{GIM}^i(D) = \sum_{s,a} n_{D}^{i}(s,a) \times \text{impact}^{\text{team}}_{i}(s,a)$$

where $D$ indicates our dataset, $\text{team}_i$ denotes the team of player $i$, and $n_{D}^{i}(s,a)$ is the number of times that player $i$ was observed to perform action $a$ at $s$. Because it is the sum of differences between subsequent Q values, the GIM metric inherits context-sensitivity from the Q function.

6.2 Rank Players with GIM

Table 4 lists the top-20 highest impacts players, with basic statistics. 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, we find that the total number of passes performed by Taylor Hall is exceptionally high at 320. 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.

7 Empirical Evaluation

We describe our comparison methods and evaluation methodology. Similar to clustering problems, there is no ground truth for the task of player evaluation. To assess a player evaluation metric, we follow previous work [Routeley and Schulte, 2015; Pettigrew, 2015] and compute its correlation with statistics that directly measure success like Goals, Assists, Points, Play Time (Section 7.2). There are two justifications for comparing with success measures. (1) These statistics are generally recognized as important measures of a player’s strength, because they indicate the player’s ability to contribute to game-changing events. So a comprehensive performance metric ought to be related to them. (2) The success measures are
Table 4: 2015-2016 Top-20 Player Impact Scores

This allows us to assess how quickly different metrics acquire predictive power for the remaining two (TOI and PIM), GIM is comparable to SI and SI show the highest correlations with success measures. EG is only the fourth best metric, because it considers only the expected value of shots without look-ahead. The traditional sports analytics metrics correlate poorly with almost all success measures. This is evidence that AI techniques that provide fine-grained expected action value estimates lead to better performance metrics. With the neural network model, GIM can handle continuous input without pre-discretization. This prevents the loss of game context information and explains why both GIM and GIM-T1 performs better than SI in most success measures. And the higher correlation of GIM compared to GIM-T1 also demonstrates the value of game history. In terms of absolute correlations, GIM achieves high values, except for the very rare events OTG, SHG, SHP and FOW. Another exception is Penalty Minutes (PIM), which interestingly, show positive correlation with all player evaluation metrics, although penalties are undesirable. We hypothesize that better players are more likely to receive penalties, because they play more often and more aggressively.

7.3 Round-by-Round Correlations: Predicting Future Performance From Past Performance

A sports season is commonly divided into rounds. In round $n$, a team or player has finished $n$ games in a season. For a given performance metric, we measure the correlation between (i) its value computed over the first $n$ rounds, and (ii) the value of the three main success measures, assists, goals, and points, computed over the entire season. This allows us to assess how quickly different metrics acquire predictive power for the full season total, so that future performance can be predicted from past performance. We also evaluate the auto-correlation of a metric’s round-by-round total with its own round total. The auto-correlation is a measure of

The model has been trained off-line, the GIM metric can be computed quickly with a single pass over the data.
We find both GIM and GIM-T1 eventually dominate the predictive value of the other metrics, which shows the advantages of modeling sports game context without pre-discretization. And possession-based GIM also dominates GIM-T1 after the first season half, which shows the value of including play history in the game context. But how quickly and how much the GIM metrics improve depends on the specific success measure. For instance, in Figure 5, GIM’s round-by-round correlation with Goal (top right graph) dominates by round 10, while others require a longer time.

### 7.4 Future Seasons: Predicting Players’ Salary

In professional sports, a team will give a comprehensive evaluation to players before deciding their contract. The more value players provide, the larger contract they will get. Accordingly, a good performance metric should show temporal consistency.

We focused on the expected value metrics EG, SI, GIM-T1 and GIM, which had the highest correlations with success in Table 5. Figure 5 shows metrics’ round-by-round correlation coefficients with assists, goals, and points. The bottom right shows the auto-correlation of a metric’s round-by-round total with its own season total. **GIM is the most stable metric** as measured by auto-correlation: after half the season, the correlation between the round-by-round GIM and the final GIM is already above 0.9.

We find both GIM and GIM-T1 eventually dominate the predictive value of the other metrics, which shows the advantages of modeling sports game context without pre-discretization. And possession-based GIM also dominates GIM-T1 after the first season half, which shows the value of including play history in the game context. But how quickly and how much the GIM metrics improve depends on the specific success measure. For instance, in Figure 5, GIM’s round-by-round correlation with Goal (top right graph) dominates by round 10, while others require a longer time.
that players get credit only for recorded individual actions. An influential approach to extend credit to all players on the rink has been based on regression [Macdonald, 2011; Thomas et al., 2013]. A promising direction for future work is to combine Q-values with regression.

Acknowledgements

This work was supported by an Engage Grant from the National Sciences and Engineering Council of Canada, and a GPU donation from NVIDIA Corporation.

References


