深層強化学習に基づくトレーディングカードゲーム環境の構築

Traiding Card Game Environments based on deep reinforcement learning

西村 昭賢 *1　森 直樹 *1　岡田 真 *1
Shouken Nishimura　Naoki Mori　Makoto Okada

*1大阪公立大学
Osaka Metropolitan University

Recently, the application of deep reinforcement learning to game environments has attracted attention. In particular, game with imperfect information has been paid attention to in this field, such as mahjong and poker. In this research, we focus on the trading card game (TCG). TCG is more difficult to be played with artificial intelligence than other games because the performance and types of available cards can be changed. This nature also makes it difficult to adjust the game balance, and it is common for the game to be modified after its release, and terms such as "buff" to change the card performance upward and "nerf" to change it downward are used. Based on the above background, we propose game balance optimization methods for TCG environments using deep reinforcement learning and evolutionary computation and demonstrate the effectiveness of the proposed methods through numerical experiments using our own TCG environment.

1. はじめに

近年, ゲーム環境への深層強化学習の応用が注目されている. 特に, プレイヤーに与えられる情報が部分的である不完全情報ゲームへの応用が注目されている. 本研究では不完全情報ゲームであるトレーディングカードゲーム (TCG) に着目した. TCG は使用可能なカードの性能や種類を変更可能という点で, 他のゲームよりも人工知能による攻略が困難である. また, この性質のためゲームバランスの調整が難しく, 公開後に修正が入ることが一般的であり, カードの性能を上方修正するバフや下方修正するナーフなどの用語が用いられる. 上記の背景から, 筆者らは深層強化学習とそれに基づく進化型計算を用いた TCG 環境のゲームバランス最適化手法を提案し, 独自の TCG 環境を用いて数値実験により提案手法の有効性を示す.

2. 要素技術

2.1 Deep Q Network

Deep Q Network (DQN) は価値ベースの強化学習手法の Q 学習と深層学習を組み合わせた代表的な深層強化学習手法である [Mnih 13]. Experience Replay などの工夫により安定した学習を可能としている.

2.2 Genetic Algorithm

Genetic Alogorithm (GA) とは, 生物の進化とその過程を模した最適化手法であり, 主に組合せ最適化問題に対して適用される. GA では解を個体として表現し, 個体群を用いて解空間の多点を同時に探索する. 各個体は良好さの指標として適応度を持つ. GA では, 個体群に対して選択, 交叉, 突然変異の 3 種類の遺伝演算子を適用し世代と呼ばれる探索を進めていく.

3. 提案手法

3.1 提案手法 1：数値実験用の TCG 環境

Magic：The Gathering *2 に代表される一般的な TCG と同様に, ゲームは 2 人のプレイヤーからなり, プレイヤーは複数のカードからなるデッキを持つ. プレイヤーは手札, 盤面と呼ばれるカードを保有する領域を持ち, ドローと呼ばれる操作でカードをデッキから手札に加える. また, プレイと呼ばれる操作でカードを手札から盤面に出す. また, デッキからカードが無くなった状態をデッキ切れと呼ぶ. また, プレイヤー自身が HP, マナという 2 つの整数値パラメータを持つ. 今回の環境ではプレイヤーの HP の最大値は 20, マナの上限値は初期値が 1 で最大値を 5 と設定した.

また TCG では, カードは場に残るユニットと使い切りのスペルの 2 種類のカードに分けられる. 本研究ではユニットのみを考慮した. カードはそれぞれ攻撃力と HP とコストの 3 つの整数値パラメータを持つ. カードがプレイされる際, プレイヤーはカードのコスト分マナを減少させる. またマナが負の値になる場合はプレイすることができない. また, 盤面にあるカードは対戦相手の盤面にあるカード, あるいは相手プレイヤーに攻撃することができる. カードが攻撃する際には, 攻撃対象の HP へとカードの持つ攻撃力分ダメージを与える. またカードへと攻撃する際には攻撃対象のカードが持つ攻撃力分, 攻撃するカードもダメージを受ける. カードが攻撃が可能となるのはプレイされたターンの次のターンからとなる. カードの HP が 0 になった, あるいは後述する手札と盤面の枚数制限を超えて手札にドロー, 盤面にプレイされた時はカードは破壊される. 破壊されたカードはゲームから取り除かれる. また, カードによっては以下に示す特殊効果を持つものもある.

召喚： プレイ時 (攻撃力 , HP) = (1 , 1) のユニットを追加でプレイ
治癒： プレイ時自プレイヤーの HP を 2 回復する
攻撃： プレイ時敵プレイヤーの HP を 2 減らす
取得： プレイ時自プレイヤーは 1 枚カードをドローする
速攻： プレイされたターンに攻撃できる

また, ゲームの流れは以下のようになっている.

1. ゲーム開始時に各プレイヤーは自身のデッキをシャッフル.
2. デッキから初期手札としてカードを 5 枚ドロー.
3. 先攻プレイヤーは 1 ターン目のドローステップをスキップし行動.
4. 後攻プレイヤーはカードを 1 枚ドローして行動.
5. 2 ターン目以降は先攻プレイヤーもカードを 1 枚ドローしてから行動.

連絡先: seb01100@st.osakafu-u.ac.jp
*2 https://magic.wizards.com

6. 4 , 5 の繰り返し．なお, ターンプレイヤーは行動前にマナを上限値まで回復．このときマナの上限値が 5 でなければ上限値を 1 増やしてから回復．
7. プレイヤーがデッキ切れになっている状態でカードをドローしようとした, あるいはプレイヤー自身の HP が 0 となった場合はそのプレイヤーが敗北となりゲーム終了．

本構築環境では一般的な TCG と同様にカードがプレイされた次のターンから行動可能となるため, 先攻プレイヤーがカードの行動が早くなり有利となる．そのため, 先攻の 1 ターン目のドローステップをスキップしている．

3.2 提案手法 2 : DQN によるデッキ内のカードパワーの定量的な評価方法

TCG においてデッキ内の各カードのカードパワーを測る指標として一般的なものはカードの HP を h, 攻撃力を a, コストを c とすると $\frac{(h+a)}{2c}$ として数値化されるマナレシオがある．しかし, マナレシオはカードの特殊効果といった要素を考慮していないためあくまで目安にしかならない．

本研究では, TCG 環境において DQN を用いた定量的なカードパワー評価指標を提案する．具体的には, DQN を用いてカードパワーを測定したいデッキにおける妥当な戦略を持つエージェントを構築する．そしてそのエージェント同士を先攻後攻両方に配置し, 先攻後攻ごとにそれぞれ 1 種類ずつカードを除いて勝率を計算する．これによりデッキ内のカードの種類数 n_{card} について, n_{card}^2 個の勝率の値を得ることができ, これらの値から定量的に構築戦略下のカードパワーを評価する．

3.3 提案手法 3 : 調整するカード枚数を最小限に抑えた TCG 環境のゲームバランス最適化手法

前提として, 本研究では TCG 環境においてデッキ間の勝率が 50% に近いことが TCG 環境におけるゲームバランスとして好ましい状態と定義する．

Fernando らは HearthStone [*1] の TCG 環境内においてデッキ間の勝率が 50 % となるように, GA を用いてデッキ内のすべてのカードのパラメータを対象としてパラメータを調整した [Mesentier Silva 19]．結果として, バランス調整の際の変更が多いと開発者やユーザーに対する影響が多いため好ましくないという考えのもと, 多目的 GA を用いることでパラメータの変更量を減らしながらデッキ間の勝率を最適化していた．しかし, 関連研究の手法ではデッキ内のすべてのカードを対象としているためどのカードにも調整が起こりえる．

そこで本研究では, TCG 環境のゲームバランス最適化の過程においてパラメータの変更量ではなく調整されるカードの枚数を抑えた方が, ゲームの開発者やユーザーにとって好ましいと考え, 調整されるカードの三数を最小限にするような TCG 環境のゲームバランス最適化手法を提案する．具体的には, 提案手法 2 によって得られたカードパワーの評価からデッキにおいて調整されるカードを限定し GA の解空間の次元を削減することで, 直接的に調整されるカードの数を減らす．

4. 実験方法

独自の TCG 環境下において以下の 3 種類の数値実験をして提案手法の有効性を確かめた．

4.1 実験 1

学習側のデッキに 3.2 節で述べたマナコストの観点において強いカードを 1 種類, 弱いカードを恣意的に 1 種類ずつ入れ, 提案する TCG 環境において DQN を適用した．後攻プレイヤーの行動のみを学習し, 学習後 10000 回対戦を実行した．学習中, 学習後の対戦において先攻プレイヤーには筆者らが実装したアグロ, コントロールと呼ばれる戦略を持つエージェントを配置している．アグロは相手プレイヤーへの攻撃を優先する戦略で, コントロールは相手盤面のカードの処理を優先する戦略である．また, 先攻プレイヤーは 1 エピソードごとに等確率で戦略を変化させ, 戦略に応じて事前に戦略間の勝率が $50 \pm 5\%$ となるよう調整されたデッキを持つ．実験から学習済エージェントの勝率を記録し, ベースラインと比較し DQN の妥当性を確かめた．また対戦した記録から選択した行動, 各カードのプレイされた回数を計測し学習序盤のエージェントと比較することで学習済みエージェントの行動を分析した．また, DQN におけるエージェントの状態空間, 行動空間は, 手札と盤面にそれぞれ 5, 9 枚の枚数上限を設けて定義した．表 2, 3 に本研究における状態空間と行動空間の定義を示す．

表 1: 学習側プレイヤーのデッキ

ID	攻撃力	HP	コスト	特殊効果	枚数
0	4	4	1	無し	2
1	2	2	2	無し	2
2	3	3	3	無し	2
3	4	3	4	無し	2
4	5	4	5	無し	2
5	2	2	2	召喚	2
6	2	3	3	召喚	2
7	1	1	1	取得	2
8	1	3	2	取得	2
9	2	1	2	速攻	2
10	3	1	3	速攻	2
11	1	2	2	攻撃	2
12	2	3	3	攻撃	2
13	1	1	1	治癒	2
14	1	1	5	治癒	2

表 2: 定義した状態空間

状態説明	次元数	最小値	最大値
各プレイヤーの HP	2	0	20
各プレイヤーの マナ	2	0	5
手札 1 ～ 9 の HP , 攻撃力, コスト, 特殊効果	36	0	5
自盤面 1 ～ 5 の HP と攻撃力	10	0	5
敵盤面 1 ～ 5 の HP と攻撃力	10	0	5
自盤面 1 ～ 5 が 攻撃可能かどうか	5	0	1
お互いのデッキの 残り枚数	2	0	30

- 1 ステップ終了後

$$r = 0.0$$

- 1 エピソード終了後

$$r = \begin{cases} 1.0 & \text{(学習プレイヤーの勝利)} \\ -1.0 & \text{(敵プレイヤーの勝利)} \end{cases}$$

また, 表 4 に実験 1 で用いた DQN のパラメータを示す．今回の実験において, ϵ-greedy における ϵ は (1) 式に従って, $\epsilon_{\max}$ から $\epsilon_{\min}$ へと指数関数的に減少する．

$$\epsilon = \max(\epsilon_{\min}\ ,\ \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min}) \exp(-\frac{n_{\mathrm{step}}}{\epsilon_{\mathrm{decay}}})) \quad (1)$$

ここで n_{step} は学習時の累積ステップ数を表す．本実験では $\epsilon_{\max} = 1.0, \epsilon_{\min} = 0.1, \epsilon_{\mathrm{decay}} = 50000$ とした．

*1 https://hearthstone.blizzard.com

表 3: 定義した行動空間

行動説明	次元数
手札 1 ～ 9 を自盤面にプレイ	9
自盤面 1 で敵盤面 1 ～ 5 に攻撃 or 敵プレイヤーに攻撃	6
自盤面 2 で敵盤面 1 ～ 5 に攻撃 or 敵プレイヤーに攻撃	6
自盤面 3 で敵盤面 1 ～ 5 に攻撃 or 敵プレイヤーに攻撃	6
自盤面 4 で敵盤面 1 ～ 5 に攻撃 or 敵プレイヤーに攻撃	6
自盤面 5 で敵盤面 1 ～ 5 に攻撃 or 敵プレイヤーに攻撃	6
ターンエンド	1

表 4: DQN のパラメータ

パラメータ名	値
割引率 γ	0.99
全結合層の活性化関数	ReLU
全結合層の次元	64
最適化アルゴリズム	Adam
方策	ϵ-greedy
Target Network 更新重み	0.5
Exprience Memory 開始ステップ数	1.0×10^5
学習ステップ数	1.0×10^6

4.2 実験 2

実験 2 では実験 1 で構築した学習済エージェントをそのまま利用する．エージェントを先攻後攻両方に配置し，表 1 のデッキを両方に持たせる．それぞれデッキからカードを 1 種類ずつ除いて 10000 回対戦を実行し先攻側の勝率を記録する．また，比較対象として構築したエージェントだけでなく筆者が作成したアグロの戦略を持つプレイヤーにおいても表 1 のデッキを持たせて先攻後攻両方に配置し，同様にカードを 1 種類ずつ除いた際の勝率を記録する．

4.3 実験 3

TCG 環境に表 1 のデッキをアグロ用のデッキとして追加し，デッキ間の勝率を 50 % に近づけるようにゲームバランスを調整するという問題を設定する．この問題設定下でカードの調整枚数を限定してデッキ間の勝率を最適化する単目的 GA を適用した．また，比較手法として 3.3 節で述べた関連研究で用いられていた表 1 の 15 種類のカードすべてを対象とした，勝率を最適化する単目的 GA および勝率とパラメータの変更量の 2 つを最適化する多目的 GA を適用した．ゲームバランスの調整において GA で調整するパラメータは表 1 の 15 種類のカードの HP, 攻撃力, コストとした．

また，単目的 GA における適応度，多目的 GA における目的関数として各解においてデッキ間の勝率に関する適応度 f_{w}，パラメータの変更量に関する適応度 f_{p}，調整されたカード種類数 f_{c} の 3 つの適応度を定義した．まず，f_{w} については，表 5 に示される問題設定において，$f_{\mathrm{w}} = \exp(-\sum_{\mathrm{i}=0}^{4} \sqrt{(0.50 - r_{\mathrm{i}})^2})$ と計算する．また，f_{p}，f_{c} に関しては，表 1 のデッキと GA の個体が表す各カードのパラメータを比較してパラメータの変更量 p, 調整されたカードの種類数 c を計算し，それぞれ $f_{\mathrm{p}} = \exp(-\frac{p}{200})$, $f_{\mathrm{c}} = \exp(-\frac{c}{15})$ と計算する．表 6 に本研究における GA のパラメータを示す．

5. 結果と考察

以下，各実験の結果とその考察を示す．

5.1 実験 1 : 結果と考察

表 7 に学習後 10000 回対戦を実行し勝率を計算した結果を示す．この時，ベースラインとして，後攻に筆者らが実装した戦略 2 種類を基に行動するプレイヤー，表 3 に沿ってランダムに行動するプレイヤーを配置し表 1 のデッキを持たせて 10000 回ゲームを実行した際の勝率も示している．これらのベースラインと比べ，学習済みエージェントは高い勝率を得ていることがわかる．深層強化学習を用いることで自動的にルールベースで作成した戦略よりも適した戦略を持つエージェントが構築できた．さらに学習済エージェントで 50000 回対戦を実行し，エージェントが構築した戦略を分析した．比較対象として，本研究において ϵ は $\epsilon_{\max} = 1.0$ から指数関数的に減少するため，表 7 でベースラインとして用いた表 3 の行動空間に沿ってランダムに行動するエージェントを学習最序盤のエージェントとして選んだ．

表 8 に 50000 回の対戦において表 3 の行動空間における各行動でエージェントが選択した総数を示す．大きな違いとしてランダムの場合はターンエンドが圧倒的に多く選ばれており，学習済の場合は 2 番目に多い行動とほぼ同数になっていた．表 3 の行動空間の定義ではエージェントが盤面の状況関係なく任意にターンエンドを選択できるようになっているが，学習を進めていくにつれて無駄なターンエンドが減っていったと分かる．また．学習済エージェントの行動に着目すると，盤面のカードで相手プレイヤーに直接攻撃するアグロ寄りの戦略を構築していることがわかる．この理由としては，表 1 の学習側に持たせるデッキにおいて強いカードとして恣意的に入れた ID 0 のカードは低コストで高い攻撃力を持つためと考えられる．

また．表 9 に 50000 回の対戦において各エージェントが表 1 内の各カードをプレイした回数を示す．表 1 から，両エージェント共にコストが小さいカードほど多くプレイされていることがわかる．ここで，恣意的にカードを追加した ID 0, 13, 7 のコスト 1 帯のカードと，ID 14, 4 のコスト 5 帯のカードに注目すると，ランダムに行動するエージェントでは同コスト帯のカードの中でそこまで顕著な差は現れていない一方で，学習済エージェントではコスト 1 帯のカードの中で ID 0 のプレイ回数が明らかに多くなっており，コスト 5 帯のカードの中で ID 14 のカードは明らかに少なくなっていた．このことから学習の結果構築戦略下におけるカードの強弱も学習していると考えられ，恣意的に入れたカードの強弱が結果に反映されており合理的であるといえる．

表 5: 実験 3 の問題設定における TCG 環境

先攻＼後攻	新追加デッキ (アグロ)	アグロ	コントロール
新追加デッキ (アグロ)	r_0	r_1	r_2
アグロ	r_3	0.5255	0.5424
コントロール	r_4	0.5121	0.5053

表 6: GA のパラメータ

パラメータ名	値
世代数	50
個体数	50
遺伝子長	調整するカードの種類数 × 3
交叉率	0.4
交叉の種類	2 点交叉
個体ごとの突然変異率	0.2
選択	1 個体だけエリート保存, その他はトーナメント方式
トーナメントサイズ	3
多目的 GA のアルゴリズム	NSGA-II

5.2 実験 2 : 結果と考察

実験 2 では，同戦略同士が表 1 を用いてカードをそれぞれ 1 種類ずつ除いた際の先攻側の勝率を計算した．よって，計 $15^2 = 225$ 個の勝率の値が結果として得られる．この 225 個の勝率の値の中で最大値，最小値に着目すると戦略下における最もカードパワーが強いカード，弱いカードを判断できる．

表 7: 後攻の戦略を変化させた場合の勝率比較

後攻の戦略	勝率
学習済エージェント	**0.7182**
アグロ	0.6914
コントロール	0.6291
表 3 の行動空間に沿ってランダム	0.2336

表 8: 選択された総数が多い行動上位 5 個

ランダム		学習済	
行動説明	総数	行動説明	総数
ターンエンド	401838	ターンエンド	245646
手札 4 を自盤面にプレイ	92053	手札 1 を自盤面にプレイ	213804
手札 1 を自盤面にプレイ	63841	自盤面 1 で相手プレイヤーに攻撃	197221
自盤面 1 で相手プレイヤーに攻撃	63458	自盤面 2 で相手プレイヤーに攻撃	103490
手札 2 を自盤面にプレイ	61304	手札 2 を自盤面にプレイ	68018

アグロ同士の対戦において最小値は先攻から ID 0 を除いて後攻から ID 14 を除いた際の先攻の勝率の値となった．このためアグロの戦略下においては，デッキにおいて ID 0 のカードが最も強いカード，ID 14 のカードが最も弱いカードと判断できる．最大値は先攻から ID 14 を除いて後攻から ID 0 を除いた場合の先攻の勝率の値であったことから同様に判断できる．

同様に，学習済エージェント同士の対戦から得られた結果と比較すると，どちらの戦略同士の対戦においても ID 0 のカードが最も強いカードと判断された一方で，最も弱いカードに関しては学習済エージェント同士の対戦結果では ID 8，アグロ同士の対戦結果では ID 14 となり，戦略によって異なる結果となった．

これは実験 1 の事前学習の効果により，学習済エージェント同士の対戦において恣意的に弱く設定したカードは登場回数が少なく勝率計算に及ぼす影響が小さかったためであると考えられる．このことから，DQN を用いて，より人間のプレイに近い戦略下におけるカードパワーを評価できることが分かった．また，学習済エージェント同士の対戦から得られた結果のように，本提案手法ではマナレシオだけでは一見判断しづらい戦略下における真に弱いカードを発見することができる可能性があるといえる．

5.3 実験 3：結果と考察

提案手法 3 の数値実験の際にはパラメータを調整するカードの優先度を決定する必要がある．本研究では，以下の手順で調整すべきカードの優先順位を決定した．

1. 得られた勝率の値の最大値，最小値を見て，最も強いカード，最も弱いカードを確かめる
2. その 2 枚において先攻後攻そのカードを除いた時の勝率を見て，カードを除いていない時の勝率と比較し，差の絶対値が大きいカードを変更する優先度が最高のカードとする
3. 先攻で そのカードを除いた状態で後攻で何も除いていない場合の勝率を計算する．
4. 先攻が優先度最高のカードを除いた場合において，3 で計算した勝率と差の絶対値を取る．
5. 値が大きいほど変更する優先度を高く設定する．

このように決定した優先順位は表 1 におけるカード ID で表すと，$0 > 8 > 6 > 10 > 4 > 1 > 3 > 14 > 2 > 11 > 13 > 9 > 7 > 12 > 5$ となった．

表 10 に，この優先順位に沿って調整するカードの種類数を増やしていった場合の結果を示す．調整するカードの種類，すなわち GA の解空間の次元が増えるほど f_w に関して良い値を持つ解が得られていた．また，調整するカード種類数が少ない個体に比べて f_w の値が小さい個体があったが，これは GA の初期収束のためと考えられる．ここで，f_w に関しては計算過程において試行ごとにばらつきがあり同一の個体で異なる値を持つ可能性がある．そのため，適応度が小さい個体においても偶然良い f_w の値を得ていた可能性があることに留意する必要がある．

表 9: 各カードが盤面にプレイされた総数 (降順)

ランダム		学習済	
ID	総数	ID	総数
13	35173	0	32965
0	35163	13	31845
7	35068	7	31581
11	31134	1	26215
9	30956	11	26184
1	30672	5	25990
5	30401	8	25986
8	30386	9	25697
12	26558	12	21763
10	26521	2	21560
6	25866	6	21393
2	25700	10	21367
3	21301	3	19382
14	18751	4	17639
4	18610	14	16807

表 10: 調整されるカード種類数を増やしながら単目的 GA を適応した結果

種類数	f_p	f_w	f_c
1	0.9048	0.4805	0.9355
2	0.8607	0.5893	0.8752
3	0.8607	0.6131	0.8187
4	0.7945	0.6339	0.7659
5	0.7334	0.6395	0.7165
6	0.6977	0.6729	0.6703
7	0.6637	0.7044	0.6271
8	0.7047	0.6837	0.5866
9	0.6505	0.7686	0.5488
10	0.6250	0.7728	0.5134
11	0.5169	0.7627	0.4803
12	0.5326	0.8078	0.4493
13	0.6005	0.8103	0.4204
14	0.5220	0.8261	0.3932

表 11: 各手法で得られた最も良好な解の適応度

手法	f_p	f_w	f_c
単目的 GA	0.44933	**0.85146**	0.36788
多目的 GA	**0.66365**	0.79097	0.42035
提案手法	0.53259	0.80783	**0.44933**

比較手法として用いたデッキ内のすべてのカードを対象とした単目的 GA，多目的 GA に関してそれぞれ最も良好な解を最終世代の適応度が最も大きい解，最終世代のパレートフロント上で目的関数 f_w の値が最も大きい解とすると，表 10 内の解で f_w に関して 2 つの比較手法の最も良好な解と中間的な解が複数得られた．その解の中で f_c の値が最も大きい解を提案手法において最も良好な解とした．

表 11 に各手法で得られた最も良好な解の適応度を示す．各手法において，それぞれの適応度に関して優越した解が得られていた．また，提案手法における最も良好な解は f_w に関して多目的 GA により得られた解に優越しており，f_c に関しては他の手法により得られた解に関して優越していた．よって調整されるカードの枚数を最小限にする TCG 環境のゲームバランス調整という提案手法 3 の有効性が確かめられた．

6. まとめと今後の課題

本研究では，深層強化学習を用いた TCG 環境の最適化手法を提案し，数値実験により有効性を示した．今後の課題として，デッキ間の相性を考慮したメタゲームの概念を取り入れた TCG 環境のゲームバランス最適化の検討などが挙げられる．

参考文献

[Mesentier Silva 19] Mesentier Silva, F. d., Canaan, R., Lee, S., Fontaine, M. C., Togelius, J., and Hoover, A. K.: Evolving the Hearthstone Meta, in *2019 IEEE Conference on Games (CoG)*, p. 1–8, IEEE Press (2019)

[Mnih 13] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. A.: Playing Atari with Deep Reinforcement Learning, *CoRR*, Vol. abs/1312.5602, (2013)