The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

強化学習と模倣学習の融合による人間らしいエージェント

Building a Human-Like Agent Based on a Hybrid of Reinforcement and Imitation Learning

ドッサ ルスラン フェルナン ジュリアン*1 Rousslan Fernand Julien Dossa
連 欣瑜 *1 Xinyu Lian
野本 洋一 *2 Hirokazu Nomoto
松原 崇 *1 Takashi Matsubara
上原 邦昭 *1 Kuniaki Uehara

*1神戸大学システム情報学研究科 Graduate School of System Informatics, Kobe University
*2株式会社エクォス・リサーチ EQUOS RESEARCH Co., Ltd.

Reinforcement learning (RL) builds an effective agent that handles tasks in complex and uncertain environments by maximizing future reward. However,the efficiency is insufficient for practical use such as game AI and autonomous driving. An effective but selfish agent conflicts with other humans,and hence the demand of a human-like behavior arises. Imitation learning (IL) has been employed to trains an agent to mimic the actions of expert behaviors provided as training data. However,IL tends to build an agent limited in performance by the expert skill,and even worse,the agent exhibits an inconsistent behavior since IL is not goal-oriented. In this paper,we propose a training scheme by mixing RL and IL for both discrete and continuous action space problems. The proposed scheme builds an agent that achieves a performance higher than an agent trained by only IL and exhibits a more human-like behavior than agents trained by RL or IL,validated by human sensitivity.

1. 序論

強化学習によるエージェントは，環境と対話しているうちに試行錯誤を重ねて学習し，様々な課題が解決できる．例えば，囲碁 [Silver 17] や自動運転 [Sallab 17, Shalev-Shwartz 16, Isele 18, Vikas 17]，テレビゲーム [Mnih 15, Mnih 16] などがあげられる．しかしながら，強化学習エージェントは収益を最大化するように訓練されるため高い性能を示すが，実用化する際には，このような性能指標以外のことも考慮する必要がある．例えば，テレビゲームにおける NPC を強化学習エージェントにすると，そのエージェントが強すぎるため，プレイヤーがゲームをあまり楽しむことができない可能性がある．また，自動運転に応用する際には，高い性能を目指して訓練された強化学習エージェントは，激しく加減速したり急に曲がったりして，隣接する車や歩行者などに不安を与える恐れがある．そこで，人間らしいエージェントを設計する必要がある．一方で，模倣学習では，人間エキスパートに提供されるデータ上でエージェントにそのエキスパートの方策を学習させるため，人間らしい態度が期待できる．ただし，学習される方策は提供されたデータに制限され，模倣学習エージェントの性能はエキスパートの性能を越えることができない．

本論文では，強化学習の高い性能を保ったまま人間らしいエージェントを設計するため，強化学習と模倣学習の融合モデルを提案する．提案した融合モデルは強化学習の高い性能と人間のような振舞を示した．実験として，離散行動空間のケースとして Atari ゲームに適用した．さらに，自動運転のような実社会にも応用可能であることを実証するために，連続行動空間の Torcs [Wymann 15] という運転シミュレータで実験を行った．評価のため，性能評価と感性試験を行い，提案モデルが人間の模倣エージェントより高い性能を示し，強化学習エージェントより人間らしく振る舞うことを実証した．

連絡先: ドッサ ルスラン，〒 657-8501 兵庫県神戸市灘区六甲台町 1-1，メールアドレス；doss@ai.cs.kobe-u.ac.jp

2. 関連研究

2.1 強化学習と Deep Q-Networks

強化学習 (RL) をマルコフ決定過程（MDP）の枠組みを通じて定義する．マルコフ決定過程は 5 つの要素 $< \mathcal{S}, \mathcal{A}, P_a, R_a, \gamma >$ からなっている．ここで $\mathcal{S}$ は状態空間であり，$\mathcal{A}$ は取りうる行動の空間である。また，$P_a(s, s') = P_r(s_{t+1} = s' | s_t = s, a_t = a)$ はあるタイムステップ t に状態 s で行動 a を取った時，状態 s' に遷移する確率であり，$R_a(s, s')$ は状態 s で行動 a を取って状態 s' に遷移すると与えられる報酬である．さらに、γ $(0 < \gamma \leq 1)$ は割引率と呼ばれ，短期報酬もしくは長期報酬のどちらを優先するかを γ で決定する．学習時，毎タイムステップ t，強化学習エージェントは $\mathcal{S}$ から状態 s を観測し，それを考慮して行動 a を取る．その結果，エージェントは環境から行動に相当する状態 s_{t+1} と報酬 r_t が与えられる．

離散行動空間の場合，Deep Q-Networks（DQN）[Mnih 15] は，様々な課題に適用され，人間より高い性能を示すことで，有効なツールであることを実証された．あるタイムステップ t，任意の状態 s からもらえる収益 $R_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$ を予測するように訓練される（T は最後のタイムステップである）．そこで，エージェントは R_t を最大化する行動 a を取る．最適な状態行動価値関数 $Q^*(s_t, a_t) = \max_\pi \mathbb{E}[R_t | s = s_t, a = a_t, \pi]$ はベルマン方程式より以下のように求められる．

$$Q^*(s_t, a_t) = \mathbb{E}_{s_{t+1} \sim \mathcal{E}}[r_t + \gamma \max_{a_t} Q^*(s_{t+1}, a_{t+1}) | s_t, a_t]$$

学習時，エージェントは収益 R_t を状態行動価値関数 $Q(\phi(s), a)$ で近似するように繰り返し更新していく．最終的に $i \to \infty$ の時，Q_i は最適な Q^* に収束する [Sutton 98].

2.2 Deep Deterministic Policy Gradient

行動空間を連続空間 $\mathcal{A} \subset \mathbb{R}^N$ とし（N は行動の次元），目標を初期分布 $J = \mathbb{E}_{r_i, s_i \sim \mathcal{E}, a_i \sim \pi}[R_1]$ の期待収益の最大化とする．DQN は高次元の状態空間上で良い性能を得られるが，低次元の行動空間に限られる．Deep Deterministic Policy Gradient（DDPG）[Lillicrap 15] は Deterministic Policy Gradient（DPG） [Silver 14] に基づく手法である．DDPG では，強化学習エージェントの方策を θ^μ でパラメータ化される actor 関

数 $\mu(s|\theta^\mu)$ とする．この関数は任意の状態 s_t から行動 a_t に写像する関数である．また，critic$Q(s_t, a_t)$ を状態と行動のペア (s_t, a_t) を近似する関数と定義する．また，actor の更新を David Silver ら [Silver 14]．に提案された通り，方策勾配により行われる．

ただし，非線形推測の critic 関数 $Q(s, a|\theta^Q)$ を導入したため，更新が発散する可能性がある．対処法として，DDPG にソフトターゲットによる更新を導入する．元々の actor と critic それぞれのパラメータ $\mu(s|\theta^\mu)$ と $Q(s, a|\theta^Q)$ を $\mu'(s|\theta^{\mu'})$ と $Q'(s, a|\theta^{Q'})$ に複製し，ターゲットネットワークと呼ぶ．そのターゲットネットワークを学習と共に $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$, $\tau << 1$ により徐々に更新していくので，学習が安定化する．

さらに，行動空間が高次元であるため，探索空間が膨大になるので，それを促進するように探索方策 $\mu'(s_t) = \mu(s_t|\theta^\mu_t) + \mathcal{N}$ を用いる [Plappert 17]．その方策には扱う課題によって異なるノイズ関数からサンプリングされるノイズを追加する．

2.3 模倣学習

模倣学習においては，エキスパートプレイヤーが従っている方策を最適な方策 π^* であると仮定し，エージェントの方策 π が π^* に近づくように学習が行われる．人間エキスパートに最適な方策 π^* により（状態，行動）ペアの系列という形で提供された上で，エージェントの方策を観測される状態からエキスパートに取れそうな行動を推測するように訓練されるので，人間エキスパートを模倣することが可能となる．

2.4 Generative Adversarial Imitation Learning

Generative Adversarial Imitation Learning (GAIL) [Ho 16] は逆強化学習と古典的な強化学習に基づく手法である．まず，逆強化学習でエキスパート方策 π_E に少ないコストを与え，他の異なる方策 π に高いコストを与えるコスト関数 c を学習する．Π は $\mathcal{S}$ から $\mathcal{A}$ に写像する方策 π の集合とし，$H(\pi) \triangleq \mathbb{E}_\pi[-log\pi(a|s)]$ を方策 π の γ で割り引かれた causal entropy [Ho 16] とする．さらに，強化学習の手法を用い，逆強化学習で求められたコスト関数 c を最小化する方策を導く．

この二段階手法の計算コストが膨大であるため，逆強化学習を用いないように J. Ho らはある方策 π の occupancy measure ρ_π を導入した [Ho 16]．ρ_π は，エージェントが方策 π に従って探索を行う際に通る (状態，行動) ペアの密度分布である．また，識別器 $D_w : \mathcal{S} \times \mathcal{A} \to (0, 1)$ を用い，エージェントとエキスパートそれぞれの方策に相当する ρ_π と ρ_{π_E} を見分ける．そのために，エージェントとエキスパートに生成される軌道による期待値の和を最小化する

$$\mathbb{E}_\pi[log(D_w(s,a))] + \mathbb{E}_{\pi_E}[log(1 - D_w(s,a))] - \lambda H(\pi), \lambda \geq 0$$

識別器 D_w がエージェントの方策とエキスパートの方策を見分けられなくなると，後者は前者に模倣されたと言える．GAIL の手続き全体はアルゴリズム 1 のようになる．

2.5 知識の蒸留

蒸留とは教師モデルの基で生徒モデルを訓練する手法である．教師あり学習に用いられるハードターゲット（one-hot label）の代わりに，教師モデルに出力させるソフトターゲット [Hinton 15] を使用する．ハードターゲットと比べて，ソフトターゲットの各要素が学習に役に立つ情報を含でいる．例えば，猫の画像を入力すると，“猫である” 確率が一番高く，“犬である” 確率は “人参である” 確率より高いと期待される．なぜならば，犬は人参よりは猫に似ていると考えられるからである．データが少量であっても良い性能が出せるだけではなく，大きなモデルを圧縮する際にも有効な手法である．蒸留が強化学習に拡張され [Rusu 15]，教師モデルから方策を導出し，より良い性能を示すということが実証されている．さらに，異なる課題に対して学習された複数のエキスパートの基で蒸留を行い，マルチタスク課題にも適用可能であることが示された．

Algorithm 1 Generative Adversarial Imitation Learning

1: **Input:** Expert trajectories $\tau_E \sim \pi_E$, initial policy and Discriminator parameters θ_0, w_0
2: **for** $i = 0, 1, 2, ...$ **do**
3: Sample trajectories $\tau_i \sim \pi_{\theta_i}$
4: Update Discriminator parameters from w_i to w_{i+1} with the gradient
5: $\hat{\mathbb{E}}_{\tau_i}[\nabla_w log(D_w(s,a))] + \hat{\mathbb{E}}_{\tau_E}[\nabla_w log(1 - D_w(s,a))]$
6: Take a policy step from θ_i to θ_{i+1}, using the TRPO rule with cost function $log(D_{w+1}(s,a))$.
7: Specifically, take a KL-constrained natural gradient step with
8: $\hat{\mathbb{E}}_{\tau_i}[\nabla_\theta log\pi_\theta(a|s)Q(s,a)] - \lambda\nabla_\theta H(\pi_\theta)$,
9: where $Q(\overline{s}, \overline{a}) = \hat{\mathbb{E}}_{\tau_i}[log(D_{w_{i+1}}(s,a))|s_0 = \overline{s}, a_0 = \overline{a}]$
10: **end for**

3. 提案手法

強化学習モデルの高い性能を保ったまま人間らしいエージェントを学習するというタスクは，(1) 人間の性能に近いエージェントを学習するサブタスクと (2) 人間のように行動を選択するエージェントを学習するサブタスクの 2 つに分割することができる．各サブタスクはそれぞれの強化学習と模倣学習の課題として取り組まれており，本論文で提案する手法はマルチタスク課題である．そこで，提案する強化学習と模倣学習の融合化手法は，離散行動空間の場合方策の蒸留に基づき，連続行動空間の場合は敵対模倣学習に基づく方法である．

また，π^* を強化学習モデルによる最適な方策，π_{HE} を人間エキスパートの方策とする．この 2 つの方策の比率を決めるパラメータを $\alpha \in (0, 1)$ とし，提案する目的関数は以下のようになる

$$\mathcal{L}_{mix}(\pi) = \alpha\mathcal{L}_{\pi_{RL}}(\pi) + (1-\alpha)\mathcal{L}_{\pi_{HE}}(\pi)$$

離散行動空間の場合，模倣学習の目的関数を，模倣学習の既存研究 [Hinton 15, Rusu 15] に従って以下の交差エントロピー損失として定義する．

$$\mathcal{L}_{\pi^*}(\pi) = \mathbb{E}_s\left[-\sum_a \pi^*(a|s)\log\pi(a|s)\right]$$

人間エキスパートの方策 π_{HE} は数理モデルとして定義するのが難しく，実験的にサンプリングされたデータの上で学習を行う．π_{HE} からソフトターゲットが得られないにも関わらず，方策の蒸留にはハードターゲットとソフトターゲットの重付き平均を計算することでより良い性能が得られる [Rusu 15]．従って，人間エキスパートに提供されるデータをハードターゲットとし，学習済み DQN モデル [Mnih 15] の方策 $\pi^{(T)}_{RL}$ の出力を熱度 T で調整したものをソフトターゲットとする．最終的に，損失関数は以下のようになる．

$$\begin{aligned}\mathcal{L}_{mix}(\pi) =& \alpha\,\mathbb{E}_{\pi_{RL}}\left[-\sum_a \pi^{(T)}_{\mathrm{RL}}(a|s)\log\pi(a|s)\right] \\ &+ (1-\alpha)\,\mathbb{E}_{\pi_{HE}}\left[-\sum_a \pi_{HE}(a|s)\log\pi(a|s)\right]\end{aligned}$$

連続行動空間の場合，模倣学習法として 2.4 節で紹介された GAIL を用いる．GAIL 法は教師モデル π からサンプリングされた軌道 $\tau \sim \pi$ を必要とする．GAIL における識別器 D_w に最大化される目的関数と生徒モデルに最小化される目的関数は

$$\mathcal{L}_{\pi^*}(\pi) = \mathbb{E}_{\tau\sim\pi}[\log(D_w(s,a))] + \mathbb{E}_{\tau^*\sim\pi^*}[\log(1-D_w(s,a))]$$

となる．ここで τ は生徒モデルからサンプリングされた軌道 $\tau \sim \pi$ である．

融合化する際には，教師モデルを人間エキスパートと強化学習モデルにするので，それぞれのエキスパートから軌道 $\tau_{HE} \sim \pi_{HE}$ と $\tau_{RL} \sim \pi_{RL}$ をサンプリングする．さらに，融合の損失関数を

$$\begin{aligned}\mathcal{L}_{mix}(\pi) =& \mathbb{E}_{\tau\sim\pi}[\log(D_w(s,a))] \\ &+ \alpha\, \mathbb{E}_{\tau_{RL}\sim\pi_{RL}}[\log(1-D_w(s,a))] \\ &+ (1-\alpha)\, \mathbb{E}_{\tau_{HE}\sim\pi_{HE}}[\log(1-D_w(s,a))]\end{aligned} \tag{1}$$

に置き換えることができる．直感的には，識別器 D_w は人間エキスパートと強化学習モデルの方策間の融合方策を認めるように学習され，この識別器を騙せるように訓練される生徒モデル π が融合方策に近づき，両方のエキスパートの長所を模倣すると期待される．

4. 実験

4.1 Atari 2600 Game: Gopher

提案手法をまず離散行動空間の Gopher という Atari 2600 システムのゲームに適用した．このゲームの目標は，農夫として地下から地上に出てくるねずみ（Gopher）が人参が取れないように，左右に動いたり穴を埋めたりすることである．人間エキスパートと訓練済みの DQN モデルがそれぞれ 55，000 のフレームを提供した上，訓練セットを 50，000，テストセットを 5，000 として学習を行った．特に，生徒モデルを訓練するために学習率 10^{-4} の Adam optimizer [Kingma 14] と Dropout 率 0.5 を利用した．蒸留の熱度を $T = 0.1$，トレードオフ係数を $\alpha = 0.93$ として実験を行った．

4.2 Torcs

Torcs [Wymann 15] は自動運転の研究で最もよく利用されるシミュレータの一つである [Lau 16] [You]．実験は，Gym Torcs 環境 [Yoshida 16] をベースにした．エージェントの観測空間は車から端までの距離，敵の車までの距離，現在の速度や加速度など，全体で 65 の連続値からなっている．行動空間は二つの要素「左右」と「加減速」からなっており，取りうる値は [-1.0,1.0] の範囲に限られる．報酬関数は走った距離にし，強化学習モデルを OpenAI Baselines [Dhariwal 17] の DDPG を基に訓練した．さらに，人間らしさが見分けられる状況が現れるように，Torcs シミュレータに停車ボットをした上，人間エキスパートに 60 秒の 220 エピソードをプレイさせ，そのデータを収集した．単なる模倣学習エージェントを，人間エキスパートのデータ上で，強化学習モデルの訓練と同じく，OpenAI [Dhariwal 17] の GAIL を用いて訓練を行った．最後に，GAIL の識別器更新を Eq. 1 で提案された通り実施し，両方のエキスパートの影響を等しくするためにトレードオフ係数 $\alpha = 0.5$ にして提案した融合モデルの学習を行った．

4.3 人間らしさの感性試験

各モデルの評価以外は，モデルの人間らしさを評価するためにダブル・ブラインドで感性試験を実施した．その試験は男性 23 人，女性 3 人の計 26 人の審査員を対象にした．年齢は 27 から 59 歳，平均年齢は 44 歳であり，株式会社エクォス・リサーチの従業員である．全員，本調査以前に本研究の内容と資料には接触がなかった．初めに，審査員に各ゲームのルールを説明し，人間に期待できる振る舞いを理解して頂けるように各ゲームの体験会を実施した．調査の内容は，各審査員，ゲーム毎に 2 本の動画（Gopher の場合 15 秒，Torcs の場合 30 秒）を提供し，人間か AI かの判断とその理由を依頼した．

5. 結果と考察

5.1 Atari 2600 Game: Gopher

性能に関して，点数の高い順にまず，強化学習モデル (DQN)，次に提案した融合モデル，最後に人間エキスパートとその模倣となった．融合モデルは，強化学習モデルに提供されたターゲットを $\alpha = 0.8$ で優先したにも関わらず，スコアの向上が 3 点しかなく，単体の強化学習モデルの点数との差が大きい．感性試験に関して，強化学習モデルはやはりあまり人間らしくないと判断されたが，提案した融合モデルは点数だけではなく人間らしさでも人間とその模倣より高いスコアを示した．そこで，融合モデルは強化学習の目標に向かった学習傾向と人間エキスパートの振る舞いを学習できた．意外にも，融合モデルは人間エキスパートより人間らしいと判断された．その理由を解明するために，審査員のコメント分析によってよく現れた感想は，“無駄な動きが少ない”，“動きが細かい，プログラム感が動きにある” また，“穴を順番に埋めようとする” であった．従って，人間のエキスパートは，特にゲームをあまりしない審査員に高い性能が期待されていなかったと考えられる．詳細は Table 1 に記載する．

5.2 Torcs

性能の評価に関して，まず，それぞれの人間エキスパート，GAIL による人間の模倣，強化学習の DDPG と融合モデルの点数を比較した．詳細は Table 2 に記載する．実験より，GAIL は訓練済みの強化学習モデルや決定論的なボットの模倣に優れているが，人間エキスパートの模倣の効率は意外に低いことが観測された．それは，人間エキスパートの方策は複雑で，基本的なニューラルネットワークでの扱いが困難だと推定される．

一方で，提案した融合モデルは例えば，強化学習モデルの高速度や人間エキスパートの曲がり方という特徴の模倣に成功した．さらに，人間エキスパートと強化学習モデルのように全体のトラックを走れるようになった．

人間であると判断された割合が低い順に，まず，強化学習モデル（DDPG）は “走るのが速すぎる” や “高速で角を曲がる” という理由で，あまり人間らしくないと判断された．意外に，人間エキスパートは人間らしくないと判断された．同じ動画に対して，審査員のコメントが多様であるが，より低い性能を示した人間の模倣エージェントが人間エキスパートより人間らしいと判断されたため，Gopher と同じく，人間エキスパートに示された性能は高いといえる．最後に，融合モデルは最も人間らしいと判断された上，強化学習モデルに近い性能を示した．

6. 結論

本論文では強化学習モデルに示される高性能をある程度保ったまま，人間エキスパートのように人間らしく動く融合モデル

Table 1: *Gopher* の結果

	点数		調査
エージェント	平均	偏差	判断率 (%)
人間	23.87	19.81	55.70
強化学習 (DQN)	**40.30**	36.81	32.69
模倣学習	23.91	23.79	**59.62**
提案手法 (RL+IL)	26.05	24.31	**59.62**

1 位と 2 位をそれぞれ太字と下線で示す

Table 2: *Torcs* の結果

	点数 ($\times 10^3$)		調査
エージェント	平均	偏差	判断率 (%)
人間	40.17	3.63	50.00
強化学習 (DDPG)	**40.45**	0.43	30.77
模倣学習	23.99	1.16	51.92
提案手法 (RL+IL)	36.63	1.32	**61.54**

1 位と 2 位をそれぞれ太字と下線で示す

を提案した．強化学習と模倣学習の最先端手法に基づいたモデル設計し，離散行動空間のタスクである Atari 2600 ゲームと連続行動空間のタスクである Torcs シミュレータに適用した．性能を評価した上，感性試験で提案モデルの人間らしさを評価した．提案モデルは，強化学習モデルによる性能向上と人間エキスパートの振る舞いの模倣に成功した．

7. 謝辞

本研究は，ＪＳＴ，未来社会創造事業，JPMJMI18B4 の支援を受けたものである．感性試験にご協力を頂きました株式会社エクォス・リサーチに深く感謝致します．本論文の校正を行ってくれた河村和紀さんに感謝致します．

参考文献

[Dhariwal 17] Dhariwal, P. e. a.: OpenAI Baselines (2017)

[Hinton 15] Hinton, G., Vinyals, O., and Dean, J.: Distilling the knowledge in a neural network, *arXiv preprint arXiv:1503.02531* (2015)

[Ho 16] Ho, J. and Ermon, S.: Generative Adversarial Imitation Learning, *CoRR*, Vol. abs/1606.03476, (2016)

[Isele 18] Isele, D. e. a.: Navigating occluded intersections with autonomous vehicles using deep reinforcement learning, in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2034–2039IEEE (2018)

[Kingma 14] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *CoRR* (2014)

[Lau 16] Lau, B.: Using Keras and Deep Deterministic Policy Gradient to play TORCS (2016)

[Lillicrap 15] Lillicrap, T. P. e. a.: Continuous control with deep reinforcement learning, *CoRR*, Vol. abs/1509.02971, (2015)

[Mnih 15] Mnih, V. e. a.: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540, p. 529 (2015)

[Mnih 16] Mnih, V. e. a.: Asynchronous methods for deep reinforcement learning, in *International conference on machine learning*, pp. 1928–1937 (2016)

[Ortega 13] Ortega, J. e. a.: Imitating human playing styles in super mario bros, *Entertainment Computing*, Vol. 4, No. 2, pp. 93–104 (2013)

[Plappert 17] Plappert, M. e. a.: Parameter Space Noise for Exploration, *CoRR*, Vol. abs/1706.01905, (2017)

[Ross 11] Ross, S., Gordon, G., and Bagnell, D.: A reduction of imitation learning and structured prediction to no-regret online learning, in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635 (2011)

[Rusu 15] Rusu, A. A. e. a.: Policy Distillation, *CoRR*, Vol. abs/1511.06295, (2015)

[Sallab 17] Sallab, A. E. e. a.: Deep reinforcement learning framework for autonomous driving, *Electronic Imaging*, Vol. 2017, No. 19, pp. 70–76 (2017)

[Shalev-Shwartz 16] Shalev-Shwartz, S., Shammah, S., and Shashua, A.: Safe, multi-agent, reinforcement learning for autonomous driving, *arXiv preprint arXiv:1610.03295* (2016)

[Silver 14] Silver, D. e. a.: Deterministic Policy Gradient Algorithms, in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. I–387–I–395, JMLR.org (2014)

[Silver 17] Silver, D. e. a.: Mastering the game of Go without human knowledge, *Nature*, Vol. 550, No. 7676, p. 354 (2017)

[Srivastava 14] Srivastava, N. e. a.: Dropout: A simple way to prevent neural networks from overfitting, *Journal of Machine Learning Research* (2014)

[Sutton 98] Sutton, R. S. and Barto, A. G. e. a.: *Reinforcement learning: An introduction*, MIT press (1998)

[Vikas 17] Vikas, B.: Deep Reinforcement Learning approach to Autonomous Navigation (2017)

[Wymann 15] Wymann, B. e. a.: TORCS: The open racing car simulator (2015)

[Yoshida 16] Yoshida, N. Y.: Gym TORCS (2016)

[You] You, Y.: TORCS for Reinforcement Learning