jh180067

大規模な強化学習技術の実証と応用

金子 知適（東京大学大学院情報学環）

概要

本研究では、近年注目されている強化学習技術の研究を進め、大規模なデータを利用することで人工知能システムの性能を向上させることを目指す。DeepMind 社の Alpha 碁の成功が示したように、各分野で人工知能システムが人間の専門家の判断力を超える時代が到来しつつあるため、AI の判断を言葉での表現する技術への期待は高い。そのためには、高性能な人工知能システムを作成する過程で行われる、1000 万から 1 億といった高次元の機械学習の過程に注目することが有力で、関連する複数のドメインでそのような超大規模な学習の実証実験を行うことで、知見を蓄積する。今年度前半は主に囲碁についての実験を行い、一定の成果を得た。

1. 共同研究に関する情報

(1) 共同研究を実施した拠点名

東京大学

(2) 共同研究分野

- □ 超大規模数値計算系応用分野
- □ ☑超大規模データ処理系応用分野
- □ 超大容量ネットワーク技術分野
- □ 超大規模情報システム関連研究分野

(3) 参加研究者の役割分担

- 代表者　金子 知適　東京大学・情報学環　強化学習の大規模化
- 副代表者 山口 和紀　東京大学・総合文化研究科科　モデルの精緻化
- 共同研究者
  - 万代 悠作 東京大学・総合文化研究科 モンテカルロ木探索
  - 万 山川 東京大学・情報学環 不完全情報ゲーム
  - 森永 雄也 東京大学・総合文化研究科 逆辞書
  - 伊部 早紀 東京大学・総合文化研究科 統計的機械翻訳

2. 研究の目的と意義

本研究では、近年注目されている強化学習技術の研究を進め、大規模なデータを利用することで人工知能システムの性能を向上させることを目指す。また将来への応用として、学習した内容を自然言語で表現する可能性について模索する。DeepMind 社の Alpha 碁の成功が示したように、各分野で人工知能システムが人間の専門家の判断力を超える時代が到来しつつあるため、AI の判断を言葉での表現する技術への期待は高い。そのためには、高性能な人工知能システムを作成する過程で行われる、1000 万から 1 億といった高次元の機械学習の過程に注目することが有力で、関連する複数のドメインでそのような超大規模な学習の実証実験を行うことで、知見を蓄積する。人工知能システムの要素技術としては深層学習が有名となっているが、本研究では強化学習の理論と未来に関する確率的な推論を行うグラフ探索の組み合わせに特に着目する。強化学習とグラフ探索の重要性を示す例と

しては、Alpha 碁において、深層学習に与える教師データに強化学習の観点で独自の工夫がなされていること、木探索を思考の基幹として深層ネットワークはそれに組み込まれる形で用いられたことなどがあげられる。強化学習は、人が教師データを与える必要がない点で応用への期待が高い一方、大量の試行錯誤を行いそれらのデータを処理するための計算機の負荷が大きいという側面を持つ。また理論を実際のドメインに適応させる際の工夫など、未知の部分も多い。そこで、本研究では、互いに関連する複数のテーマにおいて大規模な強化学習を行い、実際のデータを総合的に分析することで汎用性や頑健性の観点でモデルを強化する。遠い目標である言葉での説明のためには、実際の判断の思考記録だけでなく強化学習の過程の分析が重要と考えられる。その緒を探すために、「何か」を言葉にする技術である、統計的機械翻訳や、逆辞書の研究との関連を模索する。

近い将来に AI システムが社会の様々な場面で導入されることはほぼ間違いないが、AI が社会に受け入れられまた活用されるためには、信頼できる AI システムの作成技術を確立する必要がある。信頼を得る努力には大きく分けて2 種類があり、性能の 良いシステムを作るという点と、AI の個々の提案を検証できる情報を理解可能な形で提供する方向がある。将棋や囲碁では、前者が達成されつつあるが、後者については、AI の提案がこれまでの人間の常識を変える良い手なのか、あるいは暴走であって棄却すべきなのかを判断することは大変難しいと示されている。現状でも思考ログを提供することは可能だが、それはそのドメインの専門家が把握できるものではないため実質的に何もできていない。本研究は、前者の方向について複数のドメインで実証し発展させつつ、後者の技術への展開を模索するという点で、大きな目標に貢献するものである。大規模な強化学習については、Alpha 碁に限らず複数の成功が報告されはじめているが、囲碁や将棋、あるいはポーカーなどの不完全情報ゲームなどの複数のドメインを比較した事例はまだない。また学習で得た内容をコンピュータに説明させようという試みはいくつかなされているが、単体の予測器の範囲がほとんどで、グラフ探索と組み合わせて動作する場合の研究はない。したがって、これらについて本研究で取り組む意義がある。

3. 当拠点公募型共同研究として実施した意義

人工知能システムが高性能な判断力を備えるためには、大規模な学習が必要である。規模は年々大きくなり、一般の研究者が個人で所有する設備では難しくなっている。まず、思考ゲームの分野では 1 年前の時点で、3,000 万局の棋譜が必要であると報告されている 。棋譜は 1 手ごとのグラフ探索を行いその結果を記録したものである。このグラフ探索は、機械学習で得た評価関数 (value network)による状況判断に基づく。1 手あたり 0.1 秒で着手したとしても、それだけ棋譜を揃えるためには 10 年単位必要な計算であり、それだけでも並列分散処理や他の工夫が必須である。さらに本研究で扱う強化学習は、試行錯誤と経験からの学習を繰り返すものであり、(1)自身の対局により上記の規模の棋譜データを作成する (2) その経験から学習し自身の思考方法を調整するということを、10 から 100 回程度繰り返すだけの計算資源が必要となる。本研究には、自作のプログラムだけでなくこの分野で標準的なオープンソースソフトウェアも利用するため、Intel CPU, Linux システム, Nvidia Pascal 世代の GPU と

いった他の研究者が標準的に用いているハードウェアで実験を行う必要がある。総合して、Reedbush-{U,L}システムを活用することが適している。

4. 前年度までに得られた研究成果の概要

前年度は囲碁と将棋の研究で主に成果を得た。このうち、今年度の報告と関連が深い、囲碁についての成果の概要を述べる。この成果は、一度投稿したものをrevise中であり、RankNet for Evaluation Functions of the Game of Goとして、英文校正が完了次第、再度投稿される予定である。

この研究では東京大学も含めて国内の通常の研究機関で利用可能な計算機資源は、Google/DeepMind社よりはるかに劣るという状況に対応するために、比較的少ない教師データから知識を吸収するという点での学習効率の向上に取り組んだものである。

局面の組合せを利用した囲碁評価関数の学習法について提案した。深層学習において、複数の入力を持つネットワークは近年盛んに研究されており、様々な応用例が提案されている。そのような複数の入力を持つネットワークの学習を注意深く応用することにより、教師例の数を二乗のオーダーで増やすことが可能と期待される。

近年の人工知能の発展には、 探索手法の向上とならんで深層学習 (deep learning) 技術の進化が背景にある。深層学習によって以前は困難だと考えられてきた囲碁の評価関数の学習が成功し、 人間と同等以上の性能を発揮した。その評価関数の学習においては人間の熟達者が残した棋譜を用いて評価関数を学習し、一定の強さをもつエージェントを作成したのちに強化学習によってさらに性能のよい評価関数を学習している。

一方で、 近年様々な深層学習の応用研究がなされており、そのうち複数の入力を持つ深層ニューラルネットワーク (DNN) の研究が近年注目されている。特に入力の数が二つの DNN は Siamese ニューラルネットワークと呼ばれ、古くから研究されており、 最初期の例では署名の検証の研究において提案されている。

いずれの例でも、 入力が二つ存在するネットワークの訓練では、 N 個の教師データを組合せを用いることで O(N^2)に増やすことができ、 既存の知見が十分に蓄積されていない領域でも教師あり学習を効率的に行える可能性がある。

本研究では上記のような、 二つの入力を持つ DNN を利用し、 囲碁の評価関数を作成することを目的とした。具体的には入力として二つの局面を受け取り、 どちらがどれだけ優れているを判定する DNN の学習する。このような特徴を持つ DNN の先行研究として、ランキング学習 (learning to rank) で用いられている RankNet が挙げられる。RankNet は入力を二つ受け取る Pairwise な学習手法であり、 そのどちらが優れているかどうかを学習することが可能である。

RankNet の学習とは DNN f の学習であり，二つの入力 x_i, x_j から s_i = f(x_i), s_j = f(x_j) を計算した後、s_i, s_j と S_{ij} を用いて損失 C を計算し，誤差伝搬法によって f の重みを更新する。提案

手法は、局面の組合せを利用した RankNet を用いて囲碁評価関数の学習手法をおこなうことである。

つまり、二つの局面を受け取って、どちらの局面が優れているかを出力するネットワークの学習法について提案した。まずはじめに用意した棋譜集合を手番の勝利局面集合 W, 手番の敗北局面集合 L に分割し、それぞれの集合を訓練集合 (W_{train}, L_{train}) とテスト集合 (W_{test},L_{test}) に分割する。学習の際には、W_{¥rm train}, L_{¥rm train} からランダム一つずつ局面 w_i, l_j を抽出し、順番もランダムに入れ替えた局面対と教師例 (x_i, x_j, S_{ij}) を作成する。ここで x_i, x_j は局面の特徴ベクトルであり, どちらかが勝利局面 w_i を表す特徴ベクトルで、もう一方が敗北局面 l_j を表す特徴ベクトルである。ここで, x_i が w_i に対応するならば S_{ij} = +1, そうでないならば S_{ij} = -1 となる。その後, それぞれの入力のスコア s_i, s_j を DNN を用いて計算し, スコアから損失 C を求め, 誤差伝搬法によって DNN の重みを更新する。学習した DNN は局面の特徴ベクトルから実数値へと写像する関数 f となり、なおかつ二つの局面 s_i, s_j に関して s_i が s_j よりも優れているならば f(s_i) > f(s_j) であると期待することができる。よってこの f そのものを評価関数として利用できると期待できる。

実験では学習に用いる棋譜の数を変化させて DNN を訓練し、 正答率、 交差エントロピー損失、 そして対戦成績という側面から性能を評価し、 棋譜の数が少ない状況で, 既存手法より高い勝率を得た。実験はすべて九路盤で行った。実装には python 3 を、深層学習のフレームワークとして chainer mn を使用した。実験は RankNet を用いた DNN の学習の性能評価、学習した DNN の対戦における強さの測定、 及び着手の予測性能について行った。

提案手法で用いる DNN は、特徴ベクトルを実数値に写像するものであれば任意のものを使用できる。ここでは AlphaGo および AlphaGo Zero にそろえて実験を行った。具体的な実験結果については、出版された論文を参照されたい。

5. 今年度の研究成果の詳細

今年度の研究のうち、囲碁における成果を述べる。

囲碁においては Reedbush-L のノードを複数用いた大規模分散環境を利用し，自己対戦棋譜を用いた深層ネットワークの学習についての実験を行った．（実験は一般的に人間が対局に用いる19路盤ではなく，盤面のサイズを縮小した9路盤にて行っている．これは学習時間が19路盤に比べ 1/100 程度になるため，9路盤で知見を蓄積した後に適用することで効率的に研究を勧められるからである。教師データとして最大で約1億局面を用いて学習を行い，データや用いる知識の差異による性能の変化などを調査した．

この規模の学習を一般的なワークステー

ションで行うとすると数週間から数ヶ月かかる試算であるが，Reedbush-L を4ノード(NVIDIA Tesla P100 を合計16基使用）を用いた場合，数十時間で学習が終わるという利点がある．このため，短期間で実験のサイクルを回すことが可能になり，様々な知見を得ることができた．学習プログラムの実装には深層学習ライブラリである chainer を利用し，学習時には chainerMN と呼ばれる MPI を用いた分散実行ライブラリを用いた．

TAAI2018に採録された論文 An Alternative Multitask Training for Evaluation Functions in the Game of Go **について述べる**:

深層学習を用いた局面評価関数を作成手法を他の様々な領域へ適用するための基礎的な研究を行ったものが 本論文である．AlphaGo の後継版である AlphaGo Zero は人間の棋譜を用いることなく，自己対戦による強化学習により棋力が向上できることを示した．AlphaGo では2つの異なる目的をもつ深層ネットワークを別々に学習している．一つは value network と呼ばれる局面の価値を予測するネットワークで，もう一つの policy network はある局面での次の着手を予測する．しかし AlphaGo Zero ではその2つの深層ネットワークを統合して単一のネットワークが2つの出力をするように改良した．これにより学習も2つの性質が異なる目的をもつようになり，学習が多様化することで得られる深層ネットワークの汎化性能が向上するという知見が得られている（多様化うんぬんは適当に書きました）．この深層ネットワーク構造は囲碁だけでなくチェス・将棋でも有効であることが確かめられている．

しかしチェスや将棋のようなゲームでは着手を予測する深層ネットワーク構造は複雑になる．一般的には考えうる着手すべてを深層ネットワーク上で符号化する必要があり，チェス・将棋では組み合わせが膨大になる．そのため value network が出力するような，単純なスカラ値を予測する深層ネットワーク構造をもつ discriminative network を value network を提案されている．

研究では囲碁において policy network と value network を持つ深層ネットワークと，discriminative network と value network を持つ深層ネットワークの比較を行った．

Reedbush-L を用いて学習した際に得られた知見を活かし，深層ネットワークの構造や学習のハイパーパラメータを用いて discriminative network の学習を行った結果，policy network と value network の場合と比較して同等の性能が得られることが確かめられた．

つづいて23回ゲームプログラミングワークショップに採録された論文、囲碁ニューラルネットワークの判断根拠の可視化について述べる。

この研究では、囲碁の学習により得られた深層ネットワークについて，推論過程の可視化を目的とした研究を行った．推論過程を可視化することで人間にとってわかりやすい判断の根拠を提示することは今後の人

工知能システムの大きな研究分野である.
この研究では囲碁の深層ネットワークを Reedbush-L を用いて学習し，学習した深層ネットワークの推論についての判断根拠を可視化することを目的とした.
深層ネットワークは局面の良さ（局面価値予測）とある局面でどの行動をするべきか（方策予測）の２つの出力を持つが，それぞれについて評価実験を行った．判断根拠を可視化するためには Saliency Map と SmoothGrad という手法を適用して評価を行った．これら２つの手法は主に画像分類において効果的であるとされている．評価実験では囲碁の局面価値と方策予測を行う深層ネットワークについても既存手法が有効であるということが確かめられた.
また AlhpaGo や他の囲碁プログラムは深層ネットワーク単体だけでなく，ゲーム木探索アルゴリズムと組み合わせて次の着手を決定している．そのため，深層ネットワーク単体についての可視化だけではコンピュータの思考を理解するという観点では不十分である．そのため，ゲーム木探索アルゴリズムの一種であり，囲碁プログラムが標準的に用いているモンテカルロ木探索と呼ばれるアルゴリズムと深層ネットワークを組み合わせた場合に用いることができる判断根拠の可視化アルゴリズムを提案し，実験的に有効性を確かめた.

今年度後半は、ICGAJournal に採録された、RankNet for Evaluation Functions of the Game of Go の実験を行った。この内容は、TAAI2018 に採録された論文 An Alternative Multitask Training for Evaluation Functions in the Game of Go の手法を元に、様々な実験を拡充し、総合的に議論したものである。ノードあたり 4 GPU あることを利用して学習も対戦実験も効率的に行うことができた。

6. 今年度の進捗状況と今後の展望

研究の柱である囲碁については、前の節で紹介したように充実した研究を行うことができた。継続的に研究を続け、大規模な強化学習に関する知見を蓄積し発信することで、研究コミュニティに貢献してきたと評価できる。
研究のもう一つの柱である将棋については、将棋については昨年度までに得られたデータをもとに、モデルの精緻化を行った評価できる。その成果の一部は Computer Shogi Tournaments: World Computer Shogi Championships and Internet Shogi Server.の技術レビューの章にまとめられている。新たな計算機実験も準備中していたが、残念ながら、大規模に実施するにはいたらなかった。
その他に、現時点では小規模な実験ではあるが、チェスと将棋を同時に学習すると、それぞれを単独で学習するよりも強くなるという、興味深い成果が得られている。この成果は Multi-Task Learning of Evaluation Functions for Chess and Shogi という論文で採録されている。この同時学習の実験環境をうまく構築できれば、開発した学習モデルを大規模な強化学習に

拡大して適用することは興味深いと考えている。チェスと将棋の同時学習の実験をReedBush上で行うには、環境を整えるのに時間がかかると予想され、年度内には実施できなかった。

今後の展望としては、学際大規模情報基盤共同利用の枠組みを一旦終了して、これまでに得られた知見を整理する時間をとったうえで、計算機実験を再開する際には有料利用によるReedBushの利用に移行する予定を検討している。その際は、囲碁将棋にかぎらず他のゲームでの強化学習の知見を総合することが有力である。当研究室では以下の論文のように(ReedBushを使わない小規模な実験における)様々な強化学習に関する知見が集積しており、良い研究を実施できると考えている。

- Yuji Kanagawa, Tomoyuki Kaneko. Rogue-Gym: A New Challenge for Generalization in Reinforcement Learning.
  Arxiv abs/1904.08129 (2019)
- Taichi Nakayashiki and Tomoyuki Kaneko. Learning of Evaluation Functions via Self-Play Enhanced by Checkmate Search.
  IEEE Technologies and Applications of Artificial Intelligence, Taiwan, 2018, pp. 126--131 (DOI 10.1109/TAAI.2018.00036)
- Hanhua Zhu and Tomoyuki Kaneko. Comparison of Loss Functions for Training of Deep Neural Networks in Shogi
  IEEE Technologies and Applications of Artificial Intelligence, Taiwan, 2018, pp. 18--23 (DOI 10.1109/TAAI.2018.00014)
- Hyunwoo Oh and Tomoyuki Kaneko. Deep Recurrent Q-Network with Truncated History.
  IEEE Technologies and Applications of Artificial Intelligence, Taiwan, 2018, pp. 34--39 (DOI 10.1109/TAAI.2018.00017)
- Tianhe Wang and Tomoyuki Kaneko. Application of Deep Reinforcement Learning in Werewolf Game Agents.
  IEEE Technologies and Applications of Artificial Intelligence, Taiwan, 2018, pp. 28--33 (DOI 10.1109/TAAI.2018.00016)

7. 研究成果リスト

(1) 学術論文

- RankNet for Evaluation Functions of the Game of Go. Yusaku Mandai and Tomoyuki Kaneko. ICGA Journal
  (採録決定, 印刷中)
- Computer Shogi Tournaments: World Computer Shogi Championships and Internet Shogi Server. Tomoyuki Kaneko and Takizawa Takenobu. IEEE Transaction of Games
  (投稿中)

.

(2) 国際会議プロシーディングス

- Yusaku Mandai and Tomoyuki Kaneko. An Alternative Multitask Training for Evaluation Functions in the Game of

Go. Technologies and Applications of Artificial Intelligence 2018, pp. 132--135 (DOI 10.1109/TAAI.2018.00037)

- Shanchuan Wan and Tomoyuki Kaneko. Heterogeneous Multi-Task Learning of Evaluation Functions for Chess and Shogi, ICONIP 2018, pp. 347-358, doi: 10.1007/978-3-030-04182-3_31
- Shanchuan Wan and Tomoyuki Kaneko. Building Evaluation Functions for Chess and Shogi with Uniformity Regularization Networks, IEEE CIG 2018, pp. 70-77, doi: 10.1109/CIG.2018.8490455

(3) 国際会議発表

- Yusaku Mandai and Tomoyuki Kaneko. An Alternative Multitask Training for Evaluation Functions in the Game of Go. Technologies and Applications of Artificial Intelligenc 2018 (台湾、11月)
- Shanchuan Wan and Tomoyuki Kaneko. Heterogeneous Multi-Task Learning of Evaluation Functions for Chess and Shogi, ICONIP 2018, (カンボジア、12月)

(4) 国内会議発表

- 万代悠作, 金子知適. 囲碁ニューラルネットワークの判断根拠の可視化. 第23回ゲームプログラミングワークショップ (箱根, 11月)

(5) その他(特許,プレス発表,著書等)