畳み込みネットワークによるNo-Limit Hold'emの研究

黄 柱皓[1,a)]　金子 知適[1,b)]

概要：カードゲームの一種であるポーカーには様々なバリエーションがあるが，その中でも特に人気のある Texas Hold'em 形式のゲームについて多くの研究がなされてきた．その多くはナッシュ均衡を求めることによりゲームを解こうとするものであったが，このような手法は多くの No-Limit ゲームのように計算量が膨大になるゲームにあっては同様の適用が難しいと考えられる．近年，畳み込みニューラルネットワーク (CNN) を用いて盤面情報をパターン認識問題として学習することでポーカーをプレイするエージェントが提案されている．本稿では，この手法を No-Limit Hold'em に適用することを考え，非常に単純なヒューリスティックプレイヤー同士，また CNN によるプレイヤー同士の自己対戦によるハンドヒストリーの学習を繰り返すことで訓練されるプレイヤーについて議論する．実験では，世代が進むごとに以前の世代の弱点を学習し効果的に利用する，強化されたプレイヤーが得られた．

No-Limit Hold'em Using Convolutional Neural Networks

Juho Hwang[1,a)]　Tomoyuki Kaneko[1,b)]

Abstract: Poker is a family of card games that has many variants. There have been numerous studies on Texas Hold'em, one of the most popular of the family. A significant portion of the studies is on finding the game equilibrium, but there are difficulties in applying this approach directly on to the No-Limit variants often with considerably greater number of game states. Recently, a Convolutional Neural Network(CNN) based poker agent that tries to learn the game state as a pattern recognition problem was proposed, and this paper attempts to apply the methods to the domain of the No-Limit variant. We discuss a poker agent that studies from hand histories played by a very simple heuristic player initially, and from self-played histories of CNN-trained models. The self-trained poker models were able to effectively train from, and exploit the weaknesses of previous generations.

1. Introduction

ポーカーはカードゲームの一種であり，様々なバリエーションがある．状態数の大きさや，不完全情報ゲームであること，またそのゲームの種類の多さから，計算機にとっては難しい問題とされている．ポーカーの中でも比較的状態数が小さい 2 人対戦の Limit Hold'em (LHE) においては，Counterfactual Regret Minimization (CFR) [9] を用いて，ナッシュ均衡解 [4] が近年求められている．

一方，本研究で対象とする No-Limit Hold'em (NLHE) では 2 人対戦であっても状態数が非常に大きくなり，同じ手法を適用して厳密なナッシュを求解することは空間計算量の観点で難しい．状態数は，LHE が 3.59×10^{13} [2] に対して，NLHE は一般的に人間のプレイヤーによってよくプレイされる 100Big Blinds (BBs) [*1] の場合で 1.14×10^{35} 程度となる．CFR を計算するために必要なメモリは LHE の場合 5.23×10^{14}Bytes 程度なのに対し，100BBs の NLHE では 1.51×10^{36}Bytes と非常に膨大である．

このように膨大な状態数のゲームを解くことを考える際には，一般に抽象化を行い状態数を削減した同様の少し小

[1] 東京大学大学院 情報学環・学際情報学府
Graduate School of Interdisciplinary Information Studies, The University of Tokyo
[a)] hwang@graco.c.u-tokyo.ac.jp
[b)] kaneko@graco.c.u-tokyo.ac.jp

[*1] Blind とは，Hold'em などのゲームで最初のベットラウンドにディーラーポジションから見て最初に行動するプレイヤー 2 人に課せられる強制的なベットを指し，Big Blind は Small Blind の 2 倍の額をベットする．この Big Blind の額は，各ベットラウンドのベット額の下限を与える．

さい問題を解くことが多い．しかしこの方法は，抽象化された問題で得られた解を元の問題に適用した際に必ずしも実用的であるとは限らないという問題点がある．

また，抽象化された問題を解く際には，抽象化されたゲームで得られた戦略 A と抽象化の粒度をより細かくしたゲーム B で得られた戦略 B を比較した際に，必ずしも元のゲームで B が A より強いとは限らない [5] ことが指摘されている．更に，ゲームに対する抽象化により，仮に性質のよい解が得られたとしても，多くの場合このような抽象化は特定の一つのゲームのみに適用できるという限界をはらんでいる．

そこで近年，他分野でも成功を収めた畳み込みニューラルネットワークを用いて，データ主導のアプローチによりポーカーをプレイするエージェントが提案されている．他のプレイヤーのハンドヒストリーを用いて訓練を行い，各アクションにより得られるチップの得失を予測する．更に，このようにプレイヤーのハンドヒストリーをもとに学習したネットワーク同士でも自己対戦を行い，さらにそのヒストリーから学習することでネットワークをより発展させることができると考えられる．

2. Backgrounds

本章ではまずポーカーの種目である Hold'em と，その一種である NLHE について説明する．ポーカーは基本的に最終的な手役の強さを競う競技であるが，手札は最後の開示までは他のプレイヤーからは見られない非公開情報となっている．最後に手役を開示するより前のラウンドで相手を全員フォールドさせることができれば勝利することができるという点は，ポーカーの特徴である．

2.1 Texas Hold'em

Hold'em は，世界的に最も人気のあるポーカーの種目の一つである．Hold'em では，プレイヤーには手札 2 枚が配られ，コミュニティカードと呼ばれる全プレイヤー共通のカード 5 枚と合わせた 7 枚のうち，任意の 5 枚で役を作る．コミュニティカードは表 1 で示すようにラウンドごとに増やされ，各プレイヤはそれを参考に各ラウンドでベット（賭け金を上げる）かフォールド (降りる) かコール (相手の賭けた額に合わせる) を選択する．

ここで，コールを行うとラウンドが終了し次のラウンドに進む．最後までフォールドしなかったプレイヤーの中で，最も強い役を所持するものが勝利し，場のチップを全て回収する．通常，1 ゲームあたりの収益の期待値を最大化することがプレイヤーの目的となる．他種目ではドローという手札を引く行為があるが，このような処理は Hold'em には存在しない．

表 1 ラウンドごとのコミュニティーカードの枚数

round	Preflop	Flop	Turn	River
# cards	0	3	4	5

2.2 NLHE

計算機によるポーカープレイヤーについての研究は，その多くが LHE に関するものであった．LHE の場合と，本研究で扱う NLHE の場合における主な違いは以下の 2 点である．

- LHE では 1 回のベットラウンドごとに 4 回までのベットが可能であるが，NLHE ではこのような制限はない
- LHE では 1 回のベットの額がラウンドごとに決まっているが，NLHE では保有している上限のチップまでの任意の額のベットが可能

このような制約条件の少なさから，NLHE の場合においてはその状態数が LHE の場合と比較し膨大になる．特に，コールやフォールドの場合はその限りでないが，ベットを行う際には，その時点でベットを行うことのできるチップ数の全ての場合の数でアクションが区別されることが特徴的である．

3. Methods

本章では，まず本稿で用いる畳み込みネットワークの構成の詳細と，NLHE の学習で必要なベットアクションの学習について述べる．また，畳み込みネットワークを用いてポーカーのゲームを学習するときに必要な，ゲーム状態の表現手法についても説明する．

3.1 Convolutional Networks

本稿で説明するポーカープレイヤーは図 1 で表すような畳み込みネットワークを用いている．まずはじめに，現在のゲームの状態と行ったアクションを $33 \times 13 \times 4$ にエンコードし，これを更にゼロパディングし $33 \times 17 \times 17$ としたテンサを入力として与える．

ネットワークは現在のゲームの状態と行ったアクションに対応するチップの増減の期待値を，バイイン[*2]で除し [-1,1] の小数として出力する．

畳み込みネットワークの具体的な構成を紹介する．まず，フィルタサイズ 5×5，フィルタ数 32 の畳み込み層を 2 つ設ける．次に 2×2 で Max プーリングを行い，更にフィルタサイズ 5×5，フィルタ数 64 の畳み込み層を 2 つ設ける．もう一度 2×2 の Max プーリングを行い，最後に 1024 出力の全結合層，ドロップ率 0.5 のドロップアウト層を設けている．

損失関数は平均二乗誤差とし，最適化には Adadelta[8] を用いた．学習時の初期ハイパパラメタは，$\alpha = 1.0, \rho =$

[*2] ゲームに一度に持ち込むことのできる最大の金額．コンピュータ同士でのポーカーでは，ハンドの開始時にこの金額にリセットされ，No-Limit 種目ではゲームの大きさを定める．

$0.95, \epsilon = 10^{-8}$ としている．全ての実装は Keras[3] を用いて行われており，全ての畳み込み層において ReLU を活性化関数として用いている．

ここで，Poker-CNN[7] との共通点や，相違点を述べる．まず共通点として，フィルタサイズや各ネットワークの層 (畳み込み層，プーリング層，全結合層，ドロップアウト層) の構成はそのまま適用している．

次に相違点であるが，まず No-Limit ゲームにおいては各アクションについてその種別だけでなくアクションのチップ数が必要になることから入力のテンサにこの情報を付加している．

次に，Poker-CNN では畳み込み層のフィルタ数をそれぞれ 24 と 48 としていたのに対し，本稿では No-Limit における上記のような特徴量の多さから，これをそれぞれ 32 と 64 にしている．また本稿では最適化関数を Adadelta としている点も，Poker-CNN からの相違点である．

Poker − CNN においては，3×3 のフィルタサイズを用い，より深層な CNN を構成することも提案されていたが，本稿では上記したような 5×5 のフィルタサイズでの構成のみについて実験した．

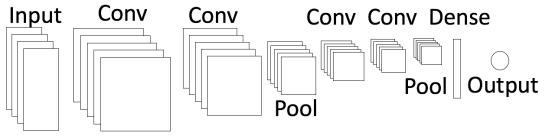

図 1 ポーカーの学習に用いた，畳み込みニューラルネットワークの構成

3.2 Learning to bet

ドローを行わない Hold'em においては，ベットについての学習のみが必要となる．また，特に NLHE の場合においては，1 回のベットラウンドにおいて選択可能なアクションの種類がとても多い．そこで，本稿では畳み込みニューラルネットワークを用い，現在の状態と行ったアクションの組み合わせに対して得られるチップの期待値を学習する．

実際のプレイングにおいては，現在の状態の情報に対して，現在行うことのできる全てのアクションを付加し，これをネットワークに入力として与える．そこで，ネットワークから予測されたチップの期待値が最も高いアクションを選択する．

まず，初期の学習については Poker-CNN に倣い，45%の確率でベット/レイズを行い，45%の確率でチェック/コールを行い，10%の確率でフォールドを行うという単純なプレイヤー (以下，ヒューリスティックプレイヤー) からシミュレーションしたハンドヒストリーを用いて学習を行う．

ハンドヒストリーから，各アクションが行われた場面のゲームの状態と，実際に行われたアクション，またそのアクションによるチップの増減を用いて，ネットワークを学習する．

ここで，本稿においては，Poker-CNN とは違いハンドの強さによってアクションの確率分布を変えておらず，常に上記の割合で一定としている点に注意されたい．

3.3 Representation

Poker-CNN において現在のゲームの状態をテンサとして表現する際に，表 2 で示すような構成でエンコードを行った．まず，2 枚の手札と，各ラウンドにおけるコミュニティーカードを 1 枚 (フロップのみ 3 枚) ずつ，それぞれ 1 枚の行列で表した．

また，手札の 2 枚と現在見えているコミュニティーカードとを全て含む全カードを 1 枚の行列で表した．次に，現在のポットサイズ[*3]を行列で表した．

たとえば，ポットサイズが 1SB であるときは行列の左上端の値が 1 となり，ポットサイズが 4SB であるときは，13×4 行列の 1 列目がすべて 1 となるといった具合である．ここで，ポットサイズが 52SB 以上である場合は，全ての要素が 1 であるような行列で表す．

ポットサイズ以外でも，アクションの金額を示すなど数値の表現が必要な場面では，すべて同様の表現でエンコードしている．また，現在のラウンド (0 から 3) や，ポジションなどの情報を示す場合には，全ての要素が 0 や 1 であるような行列を用いて表している．

現在のラウンドや，過去のラウンドに行われたアクションの履歴は，前述の通り NLHE では可変長になりうるが，本稿では Poker-CNN の論文における，2-7 トリプルドローポーカーでの例に倣い現在のラウンドでの 5 アクションと一つ前のラウンドの 5 アクションを含めた．ここで，2 つ前以上のラウンドのヒストリーは表現されない点に注意されたい．

表 2 NLHE の盤面の入力として与えたテンサの構成

特徴量	行列数	説明
手札	2	手札を 1 枚ずつ用いる
コミュニティー	3	ラウンドごとに 1 枚ずつ用いる
全カード	1	現在見えている全てのカードを示す
残余ラウンド	3	0 から 3 までの範囲を示す
ポジション	1	ラウンド開始時の行動順を示す
ヒストリー	5	現在のラウンドの過去 5 アクション
	5	上記アクションのチップ数
	5	前ラウンドの過去 5 アクション
	5	上記アクションのチップ数
アクション	1	現在のアクションの種類
	1	上記アクションのチップ数

4. Experiments

本章では，畳み込みニューラルネットワークによるポー

[*3] 現在までにベットされたチップの総和

カープレイヤーのモデルの学習について詳述し，またこのようにして作られたプレイヤーについての評価を行う．

4.1 Learning the model

まずはじめに，上述したヒューリスティックプレイヤー同士によるハンドヒストリーを 200,000 個生成した．これらのハンドヒストリーから，各アクションを抽出した．この初期の自己対戦においては，1 ハンドで平均して約 7 回アクションが行われ，約 1,400,000 個のアクションが得られた．

次に，このアクションから学習したネットワークをもとにしたプレイヤー同士の自己対戦と，ヒューリスティックプレイヤーとの対戦を用いて再びネットワークを学習する． ここで，現在のネットワークと過去のネットワーク同士の全ての組み合わせでの対戦を均等な割合のハンド数で学習させることで，特定の世代に対して過学習することを防止している．

4.2 Evaluation

本稿で述べた CNN によるポーカープレイヤーを評価するために，以下のように実験を行った．

- ランダムプレイヤー (選択可能な全てのアクションを同確率で行う)
- ヒューリスティックプレイヤー (上記，ハンドヒストリーを生成するために用いた)
- 第 1・2・3 世代の Poker-CNN プレイヤー

このような 5 タイプのプレイヤー同士での対戦 (ここで，持ちチップは 50BBs としている) を 100,000 ハンド行い，1 試合あたりの平均損益を計算した結果は表 3 に示す通りである．

表 3 上記プレイヤー同士で 100,000 ハンド対戦したときの結果 (単位は Big Blinds)

model	Rand	Heu	CNN1	CNN2	CNN3
Random	0	-6.82	0.78	0.10	-0.58
Heuristic	6.82	0	3.57	3.74	7.86
CNN-1g	-0.78	-3.57	0	-0.28	-2.35
CNN-2g	-0.10	-3.74	0.28	0	-5.32
CNN-3g	0.58	-7.86	2.35	5.32	0

4.3 Results

最初に，ランダムプレイヤーに着目すれば，初期の CNN プレイヤーはランダムプレイヤーに対しても勝ち越せていないのに対して，3 世代目からランダムプレイヤーに勝ち越せるようになっている点がわかる．

また，CNN プレイヤー同士での対戦結果からも，それぞれ自身より以前の世代に対して勝ち越せていることから，過去のハンドヒストリーから学習ができていそうであるとわかる．

その反面，全ての CNN プレイヤーがヒューリスティックプレイヤーに負け越している点や，特に世代を追うごとにヒューリスティックプレイヤーに対する負け額が増えている傾向も見てとれる．

図 2 に示すのは，各プレイヤーにおけるベットアクションの割合であるが，特に CNN プレイヤー同士で比較すると，世代を追うごとに特にフォールドの割合が少なくなっているのがわかり，これは以前の世代の弱点を学習している結果であると考えられる．

図 3 に示すのは，各プレイヤーにおけるベットサイズの割合である．ヒューリスティックプレイヤーと比べて，CNN プレイヤーにおいてはいずれも非常に大きいベットサイズを取っていることが多いため，ヒューリスティックプレイヤーに対する大きな弱点となっていると考えられる．

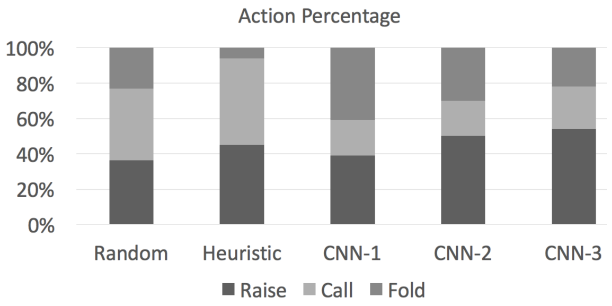

図 2 各プレイヤーにおけるベットアクションの割合

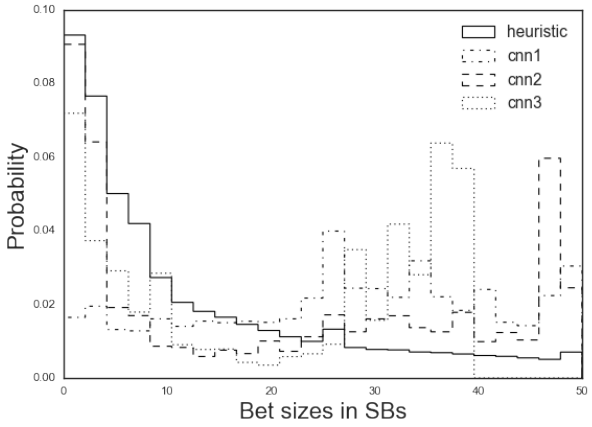

図 3 各プレイヤーにおけるベットサイズの割合

5. Related Works

コンピュータープレイヤーによるポーカーについては多くの研究がなされている．Bowling[4] らのグループによる CFR を用いた手法が代表的な例で，抽象化を施すことなくすべてのゲーム状態を探索し実質的なナッシュ戦略を得ている．

しかしながら，CFR による方法を 2 人対戦 LHE 以外の一般的な他種目，特に今回扱ったような NLHE など No-Limit の諸種目においても抽象化を行うことなく同様に適用することは，今の段階では空間計算量の問題から困難と考えられる．

また，抽象化を施すなどして元のゲームより状態数を削減したゲームに対して CFR を適用することが考えられるが，必ずしもその抽象化されたゲームで得られた解をもって元のゲームで実用的な解が得られるとは限らないし，このような方法は特定のゲーム種目に対する事前知識を要求するため汎用性が低いと言える．

Yakovenko らによる Poker-CNN においては，ポーカーのハンドをパターン認識の問題として解けると仮定することによってデータ主導でのポーカーの学習モデルを提案している．特に，どのようなゲームに対しても事前知識を要さずある程度同様のシステムによって学習できそうであるという点で，CNN を用いたデータ主導の方法は汎用性が高いと考えられる．

しかし，Poker-CNN においては，ビデオポーカー，2-7 トリプルドローポーカー，LHE という 3 種目についてのみ実験しており，NLHE など状態数が非常に膨大になる No-Limit ゲームについては論じられなかった．本稿では，畳み込みネットワークを用いた No-Limit ゲームの学習を行う際に必要となる，ゲームにおける盤面情報のエンコーディングや，フィルタ数，また学習するハンド数などについて論じた．

6. Discussion and Future Works

本稿では，既存の Poker-CNN を拡張する形で NLHE への適用を試みた．本稿での手法は，3 代目の時点ではまだ初期の棋譜を生成するのに用いたヒューリスティックプレイヤーに勝利できる強さには至らなかったが，ランダムプレイヤーに対する成績が順調に向上していたことや，CNN プレイヤー同士での対戦でもより学習の進んだ世代のネットワークが勝ち越すことができていた．

Poker-CNN でベンチマークに用いられていた CNN プレイヤーは，LHE の場合で第 8 世代，2-7 トリプルドローの場合で第 20 世代であったから，本手法による学習においても同様に計算時間をかければ，一定の強さは期待できると考えられる．

特に，図 2 で示したベットアクションの割合の推移からも，世代が進むごとにフォールドの割合が減っているので，3 世代以降学習を続けていった場合，ヒューリスティックプレイヤーに近い分布か，もしくはヒューリスティックプレイヤーより若干ベットの割合の高いプレイヤーが得られると期待される．

また，CNN プレイヤー全般においてヒューリスティックプレイヤーに対する弱点として考えられる，ベットサイズの大きさという面においても，図 3 で示すように世代が進むごとに少しずつベットサイズが小さくなっていることから，学習が進めばより強いプレイヤーが得られると期待される．

6.1 Future Works

今後の課題として，計算時間をかける以外にも CNN プレイヤーの質を向上させるためにいくつか考えられる手法が存在する．まず，Hold'em や Omaha など，多くの種目においてはカードのスート (スペード，ハートなど) は全てその手役としての強さが同等であるから，手札やコミュニティーカードに対して正準変換 [6] を施すことにより，カードに対する情報を大幅に削減[*4]できる．

次に，本稿で行った実験においては，50BBs の NLHE についての学習を行っていたが，ゲームの大きさを 20BBs などある程度制限することによって，CNN プレイヤーの質を向上させることができると考えられる．また，今回の論文では 5×5 のフィルタサイズでの畳み込みネットワークのみを検討したが，3×3 のフィルタサイズでの畳み込みネットワークについても実験を行い各ネットワークの NLHE でのパフォーマンスを比較するなどが，今後の課題として残る．

更に，ネットワークに学習させるためのヒューリスティックプレイヤーのアクションの生成について，Poker-CNN では手札の強さを考慮した確率分布によって生成を行っているが，本稿における実験では前述の通りこのような処理を行っていないことから，CNN プレイヤーのプレイングには，強いハンドであってもフォールドしたり，弱いハンドであってもレイズするなどの場面が多く見られた．

本稿では，ランダムプレイヤーやヒューリスティックプレイヤー，各世代の CNN プレイヤー同士のみでの比較を行ったが，NLHE ゲームでの CFR の実装によるプレイヤーについても強さのベンチマークとして比較を行うことが今後の課題として残る．

参考文献

[1] Brown, N. and Sandholm, T., Regret Transfer and Parameter Optimization. In *AAAI 2014*.

[2] Johanson, M, Measuring the Size of Large No-Limit Poker Games. In *University of Alberta, Technical report 2013*.

[3] Chollet, François, Keras. From *Github*.

[4] Tammelin, O. et al., Solving Heads-up Limit Texas Hold'em. In *IJCAI15*.

[5] Waugh, K. et al., Abstraction Pathologies in Extensive Games. In *AAAI Workshop on Computer Poker and Imperfect Information, 2013*.

[6] Waugh, K., A Fast and Optimal Hand Isomorphism Al-

[*4] 例えば，Preflop 時の手札の組み合わせは，${}_{51}C_2 = 1326$ 通りであるが，正準変換されたあとの Preflop の手札の組み合わせは，$13 \times 13 = 169$ 通りである

gorithm. In *Proc. of AAMAS 2009*.
[7] Yakovenko, N. et al., Poker-CNN: A Pattern Learning Strategy for Making Draws and Bets in Poker Games Using Convolutional Networks. In *AAAI 2016*.
[8] Zeiler, Matthew D., ADADELTA: An Adaptive Learning Rate Method. From *CoRR*.
[9] Zinkevich, M. et al., Regret Minimization in Games with Incomplete Information. In *NIPS 07*.