ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 19-04-11
72

j.inst.Korean.electr.electron.eng.Vol.23,No.4,1187~1194,December 2019
http://dx.doi.org/10.7471/ikeee.2019.23.4.1187

CNN 기반 기보학습 및 강화학습을 이용한 인공지능 게임 에이전트

An Artificial Intelligence Game Agent Using CNN Based Records Learning and Reinforcement Learning

전 영 진, 조 영 완*★

Youngjin Jeon, Youngwan Cho*★

Abstract

This paper proposes a CNN architecture as value function network of an artificial intelligence Othello game agent and its learning scheme using reinforcement learning algorithm. We propose an approach to construct the value function network by using CNN to learn the records of professional players' real game and an approach to enhance the network parameter by learning from self-play using reinforcement learning algorithm. The performance of value function network CNN was compared with existing ANN by letting two agents using each network to play games each other. As a result, the winning rate of the CNN agent was 69.7% and 72.1% as black and white, respectively. In addition, as a result of applying the reinforcement learning, the performance of the agent was improved by showing 100% and 78% winning rate, respectively, compared with the network-based agent without the reinforcement learning.

요 약

본 논문에서는 인공지능 오델로 게임 에이전트를 구현하기 위해 실제 프로기사들의 기보를 CNN으로 학습시키고 이를 상태의 형세 판단을 위한 근거로 삼아 최소최대탐색을 이용해 현 상태에서 최적의 수를 찾는 의사결정구조를 사용하고 이를 발전시키고자 강화학습 이론을 이용한 자가대국 학습방법을 제안하여 적용하였다. 본 논문에서 제안하는 구현 방법은 기보학습의 성능 평가 차원에서 가치평가를 위한 네트워크로서 기존의 ANN을 사용한 방법과 대국을 통한 방법으로 비교하였으며, 대국 결과 흑일 때 69.7%, 백일 때 72.1%의 승률을 나타내었다. 또한 본 논문에서 제안하는 강화학습 적용 결과 네크워크의 성능을 강화학습을 적용하지 않은 ANN 및 CNN 가치평가 네트워크 기반 에이전트와 비교한 결과 각각 100%, 78% 승률을 나타내어 성능이 개선됨을 확인할 수 있었다.

Key words : Reinforcement Learning, Othello game agent, Value function network, CNN, Records learning

Dept. of Computer Engineering, Seokyeong University
★Corresponding author
E-mail : ywcho@skuniv.ac.kr, Tel : +82-2-940-7749

※ Acknowledgment
This research was supported by Seokyeong University in 2019.
Manuscript received Nov. 20, 2019; revised Dec. 13,, 2019; accepted Dec. 18, 2019.

Ⅰ. 서론

최근 복잡하고 어려운 문제들을 딥러닝(Deep Learning)을 사용해 해결하려는 연구들이 활발히 이루어지고 있다. 딥 러닝은 영상 인식, 음성 인식, 자연어 처리, 빅 데이터 분석 등 다양한 분야에 활용되고 있으며 인간조차 하기 어려운 일을 대체해 나가고 있다.

딥러닝을 사용한 인공지능 중 바둑과 같이 복잡하고 상황판단을 필요로 하는 지능적 보드게임에 성공적으로 적용된 대표적인 사례로 알파고(AlphaGo)[2][3]가 있다. 2016년 프로기사 이세돌을 상대로 4:1로 승리한 알파고는 합성곱신경망(Convolutional Neural Network)과 강화학습을 사용해 구축한 의사결정과정을 통해 상황에 따른 최적의 수를 찾도록 설계되어 있다. 이후 알파고는 여러 번의 개선을 통해 AlphaGo-Zero로 발전하였으며 AlphaGo-Lee와의 대국에서 100%의 승률을 보이며 딥러닝 및 강화학습이 적용된 인공지능의 발전가능성을 보여주었다.

바둑뿐만 아닌 다른 지능적 보드게임인 체스(Deep Blue)[4], 오델로(Logistello)[5] 등에도 인공지능이 인간인 세계챔피언에게 도전하여 승리한 바가 있다. 그 중 오델로는 인공지능 연구가 활발히 이루어지는 게임 중 하나로써, 약 10^{54} 가지의 경우의 수를 가지고 있어 다른 지능적 보드게임 못지않은 많은 상태 공간을 가지고 있다. 또한 아직까지 완전한 솔루션이 제시되지 않았으며, 매년 새로운 인공지능 알고리즘들이 기성 알고리즘을 상대로 승리하며 발전해나가고 있다.

대표적인 오델로 알고리즘으로는 IAGO[6], BILL[7], Logistello, Zebra[8] 등이 있으며, 모두 가치기반의 평가함수를 사용하며 우수한 성능을 보여주었다. 그러나 모두 테이블 형태의 가치평가함수를 가지고 있어 대략 10^{54} 개의 상태공간을 가지고 있는 오델로의 모든 상태에 대한 가치평가가 어려운 문제점이 있다. 이러한 문제점을 해결하기 위해서 가치평가함수를 테이블로 구성하지 않고 인공신경망을 활용하여 근사하거나[8][9], 강화학습을 적용하여 의사결정을 하려는 연구 또한 활발히 진행되었다[10][11].

본 논문에서는 지도학습을 이용하여 실제 프로기사들의 대국을 기록한 기보를 상태 별로 분해하고, 승패에 따른 평가를 누적한 학습 데이터를 생성하여 CNN을 이용해 표현함으로써 모든 경우의 수에 따른 가치평가함수를 근사화하는 방법을 사용한다. 지도학습으로 근사화된 가치평가함수를 기반으로 최소최대 탐색(Minimax search)을 적용하여 현재의 상태로부터 진행될 수 있는 여러가지 상태들을 예측하고, 그 상태에 대한 유・불리 평가를 시행하며 이 평가를 기반으로 최선의 수를 탐색하여 최종적인 의사결정을 수행한다. 나아가 기보 학습을 통해 구축한 가치평가 네크워크 CNN을 스스로 발전시킬 수 있도록 강화학습 이론의 정책 이터레이션을 적용한 자가대국 방법을 제안한다.

Ⅱ. 게임 에이전트의 구성

본 논문에서는 오델로 게임의 각 예측된 진행 상황에서 승패의 유・불리 평가를 통한 최선의 수 선택을 위한 의사결정을 위해 프로 기사들의 기보를 학습한 가치평가 네트워크를 기반으로 최대최소탐색을 통해 주어진 상황에서 최적의 수를 결정하는 구조를 이용한다[1].

1. 심층신경망 기반 의사결정

본 논문에서는 게임에이전트의 의사결정을 위해 가치평가함수를 근사화한 네트워크를 기반으로 하여 최대최소탐색[14]을 통해 최선의 수를 선택한다.

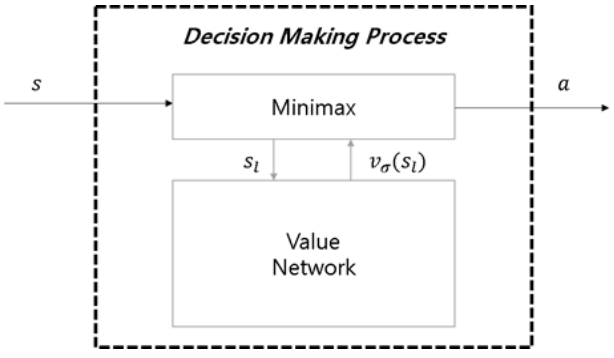

Fig. 1. Decision making structure of game agent.
그림 1. 게임 에이전트의 의사결정 구조

그림 1과 같이 게임의 특정 진행 상태 s에서 플레이어의 의사결정을 위하여 최소최대탐색의 말단 노드(leaf node) s_l에 대해 학습된 심층신경망의 출력인 $v(s_l)$을 적용하여 탐색 결정을 통해 의사결정을 수행한다. 본 논문에서 최소최대탐색과 근사화한

가치평가 네트워크를 통한 의사결정과정 $\pi_\theta(s,a)$는 식 (1)과 같이 정의한다.

$$\pi_\theta(s,a) \doteq minimax(s,a,\theta) \quad (1)$$

식 (1)에서 s는 플레이어가 의사를 결정할 현재의 게임 상태를 의미하고 a는 의사결정과정에 의해 결정된 행동을 의미하며 θ는 심층신경망의 학습 파라미터를 나타낸다.

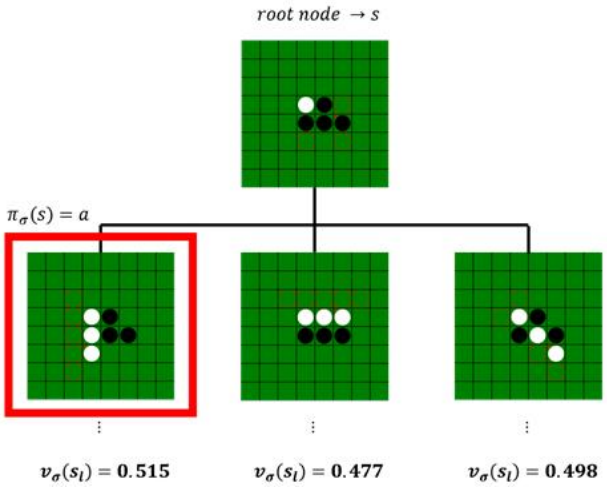

Fig. 2. Example of game agent decision making.
그림 2. 게임 에이전트의 의사결정 예시

그림 2는 본 논문에서 사용한 의사결정의 예를 간략히 나타낸 것이다. 루트 노드는 초기상태에서 *F*5지점에 착수한 상태이며, 루트 노드에서 파생하는 상태는 *D*6, *F*4, *F*6에 착수한 3가지 상태이다. 이처럼 착수 가능한 지점에 대해 파생하여 자식 노드들을 생성하게 되고, 그 생성된 자식 노드들이 지정한 깊이(depth)에 도달하게 되면 그 리프 노드들에서의 가치평가를 시행해 부모 노드로 최소, 최대를 따져가며 반환하게 된다. 이러한 과정을 통해 반환된 가치평가들이 루트의 자식 노드들에서 최대가 되는 점을 착수점으로 선택하게 된다. 본 논문에서는 이와 같이 루트 노드에서 시작하여 정해진 깊이(depth)까지 진행했을 때 가능한 경우의 예측 상태들(s_l)에 대해 학습된 심층신경망을 통해 평가된 승패 예측 확률을 근거로 최대최소탐색의 역과정을 거쳐 의사결정을 진행하며, 말단 노드의 상태 s_l에 대한 승패의 유·불리 예측(형세) 판단을 위해 프로기사들의 실제 기보를 CNN을 통해 학습하는 방법을 제안하고, 나아가 강화학습을 이용하여 CNN의 파라미터를 갱신하는 방법을 제안한다.

2. 심층신경망의 구조 및 기보학습

본 논문에서 제안하여 사용하는 심층신경망은 예측된 특정 게임 상황이 주어졌을 경우 이에 대한 형세를 평가하여 제공하는 역할을 하므로 게임의 특정 상황을 입력으로 하고 형세를 나타내는 척도로서 승패의 유·불리에 대한 확률을 학습결과로 출력하는 구조를 갖는다. 이를 위해 본 논문에서는 게임이 진행되는 각 상황을 상태 s로 정의하고 상태 s에 대한 승패의 유·불리 정도를 가치함수 $v(s)$로 정의하여 사용한다.

본 논문에서는 가치평가함수를 근사화하기 위해 그림 3과 같은 구조의 CNN(Con-volutional Neural Network)을 사용하였는데 이는 8×8의 공간을 가지는 오델로 게임의 공간적인 특성과 국지적인 형태에 따른 특징들을 추출할 수 있도록 하기 위함이다.

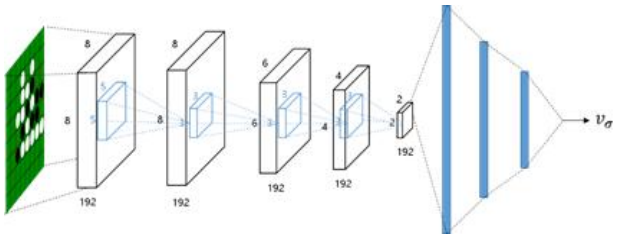

Fig. 3. Structure of CNN for state value evaluation.
그림 3. 상태 가치 평가를 위한 CNN 구조

본 논문에서는 가치평가 네트워크 CNN의 효과적 기보학습을 위해 CNN의 입력 요소로서 게임의 진행 상태뿐만 아니라 착수 차례, 착수 가능 지점, 전략적 상황의 표현으로서 이동성을 포함하였다. 이동성은 상대방의 착수 가능한 경우의 수를 줄이고, 자신의 착수 가능한 경우의 수를 극대화하려는 전략으로서 오델로 게임의 상태를 세부적으로 표현할 수 있는 전략적 특징 요소 중 하나이다.

본 논문에서 사용한 CNN의 입력 요소는 돌의 배치를 나타내기 위해 3개 채널, 착수 가능 지점을 표현하기 위해 1개 채널, 이동성을 표현하기 위해 8개 채널, 착수 차례를 표현하기 위해 1개 채널, 총 13개의 8×8 binary feature 채널로 구성되어 있다. 이동성의 경우 0~7의 해상도에 대해 각각 8×8 영역을 이진화하여 8개의 채널로 나타낼 수 있다.

Conv.1과 Conv.2층은 zero-padding을 사용하였고 출력층을 제외한 나머지 층들은 활성화 함수로 ReLU(Rectified Linear Unit)를 사용하였으며 출력층은 승패에 대한 확률로 표현되기 때문에 0~1 사

이의 값을 가지는 Sigmoid를 활성화 함수로 사용하였다.

그림 4는 기보를 이용해 생성한 최종적인 입력의 구성 및 학습데이터의 생성과정을 나타낸 것이다. 본 논문에서는 약 15만개의 프로기사들의 실제 기보 데이터를 시뮬레이션과정을 통해 분해 및 결합하여 약 730만개의 입력데이터인 상태 s들로 구성된 상태집합 S를 생성하여 사용하였다.

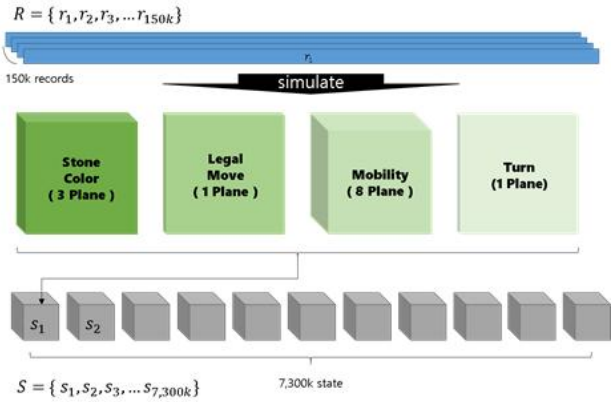

Fig. 4. Data generation process for learning records.
그림 4. 기보 학습 데이터의 생성 과정

상태별로 분류된 각각의 학습데이터를 사용하여 상태 s에 대한 형세를 나타내는 가치평가함수 $v_T(s)$는 다음 식 (2)와 (3)과 같이 정의한다[1].

$$z_k(s)=\begin{cases}1, win\\0, lose\end{cases} \tag{2}$$

$$v_T(s)=\frac{1}{n}\sum_{k=1}^{n}z_k(s) \tag{3}$$

식 (2)에서 $z_k(s)$는 상태 s가 나온 k번째 게임의 승패를 판단하며 게임에서 이기는 경우 1을, 지는 경우 0을 반환한다. 식 (3)에서 n은 상태 s가 기보에서 등장한 총 횟수를 나타낸다. 따라서 $v_T(s)$는 기보에서 등장한 특정 상태 s에 대해 승리한 경우의 확률을 나타내므로 이를 승패의 유·불리 즉, 형세를 판단하는 척도로 사용하였다.

본 논문에서 CNN의 학습에 사용하는 비용함수(Cost function)는 식 (4)와 같이 구성하였다.

$$C(\theta)=\frac{1}{2M}\sum_{s}\frac{n(s)}{N}(v_T(s)-v_\theta(s))^2 \tag{4}$$

전체 학습데이터 개수 N에 대한 상태 s의 상대적 출현 빈도인 $n(s)$를 가중한 변형된 MSE를 사용하였고 이를 통해 가치평가 네트워크의 출력 $v_\theta(s)$가 기보를 통해 구성한 가치평가함수 테이블 $v_T(s)$를 학습할 수 있도록 하였다.

III. 게임 에이전트의 강화학습

본 논문에서는 최소최대탐색과 가치평가함수를 기반으로 한 의사결정 과정을 강화학습의 정책으로 사용하여 이를 학습하는 방법을 제안한다. 게임의 어떤 진행 상태 s에 대한 가치평가 네크워크를 기반으로 탐색을 통해 다음의 수를 결정하므로 가치평가 네크워크는 강화학습의 정책을 결정하는 근거가 된다. 따라서 본 논문에서는 정책을 결정의 주요 근거가 되는 가치평가 네크워크를 자가대국을 통해 갱신하는 방법을 제안한다.

그림 5는 강화학습을 통해 가치평가 네크워크의 파라미터를 갱신하는 과정의 개요를 나타낸 것으로 각각 구조는 동일하나 파라미터가 다른 가치평가 네크워크를 사용하는 두 게임 에이전트가 대국을 실시하고, 이 과정에서 생성되는 상태 및 승부결과를 다음 단계의 파라미터 갱신에 학습데이터로 사용한다.

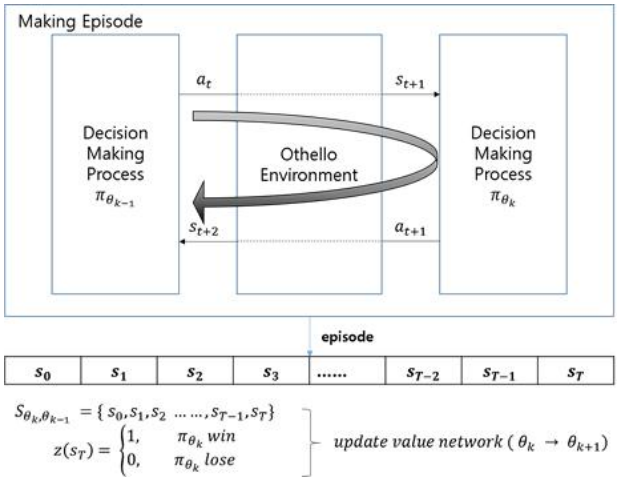

Fig. 5. Data generation process for learning records.
그림 5. 기보 학습 데이터의 생성 과정

대국을 실시하는 두 게임 에이전트는 각각 현재의 파라미터 θ_k를 가지는 네크워크 기반 정책과 π_{θ_k}와 이전 단계의 파라미터 θ_{k-1}를 가지는 네크워크 기반 정책 $\pi_{\theta_{k-1}}$을 사용하며 이를 수식으로 표현하

면 다음 식 (5) 및 (6)과 같다.

$$\pi_{\theta_k}(s) = minimax(s, depth, \theta_k) \tag{5}$$

$$\pi_{\theta_{k-1}}(s) = minimax(s, depth, \theta_{k-1}) \tag{6}$$

본 논문에서는 두 에이전트가 대국을 실시하여 게임이 종료될 때까지를 하나의 에피소드로 정의하고 하나의 에피소드에서 발생한 상태의 집합 $S_{\theta_k,\theta_{k-1}}$를 다음 식 (7)과 같이 정의한다.

$$S_{\theta_k,\theta_{k-1}} = \{s_0, s_1, s_2, \cdots, s_{T-1}, s_T\} \tag{7}$$

또한, 에피소드 종료 시점 T에서의 상태 s_T에 대한 승패의 결과 $z(s_T)$는 식 (2)와 같이 정의되며, 현재의 네크워크가 게임 도중 발생한 상태에 대해 추종하여야 할 목표치로 사용하기 위해 하나의 에피소드에서 $S_{\theta_k,\theta_{k-1}}$에 속한 모든 상태들에 대해 $z(s_T)$를 부여하여 학습데이터를 구성하고 이에 대한 학습을 위한 비용함수는 다음 식 (8)과 같이 구성하였다.

$$C(\theta_k) = \frac{1}{2}\sum_{i}^{T}(z(s_T) - v_{\theta_k}(s_i))^2 \tag{8}$$

본 논문에서는 가치평가 네크워크의 파라미터 갱신을 위해 식 (8)을 비용함수로 사용하여 경사하강법을 적용하였으며 파라미터 갱신을 위한 식은 다음 식 (9)와 같다.

$$\begin{aligned}\theta_{k+1} &= \theta_k + \alpha\frac{\partial C}{\partial \theta_k} \\ &= \theta_k - \alpha\sum_{i}^{T}\left(z(s_T) - v_{\theta_k}(s_i)\right)\frac{\partial v_{\theta_k}}{\partial \theta_k}\end{aligned} \tag{9}$$

Ⅳ. 실험 및 결과

1. 실험 환경

본 논문에서 실험에 사용한 가치함수 네트워크의 구성은 Windows 10환경에서 tensorflow와 keras를 이용해 Python으로 구현하였고, 오델로 환경과 최소최대탐색은 C#으로 구현하였다. 모든 학습과정은 1개의 GPU(Nvidia Geforce 1080 Ti 11Gb)와 1개의 CPU(Intel i7 7700k)를 통해 이루어지도록 구성하였다.

지도학습과 강화학습에 사용된 하이퍼 파라미터는 표 1에 주어진 바와 같이 두 학습과정 모두 Adam optimizer를 사용하였으며 학습률은 0.001로 고정하였고 epoch는 지도학습은 300, 강화학습은 10으로 하였다. Batch 크기는 지도학습은 1024, 강화학습은 하나의 에피소드에서 모은 데이터들을 full-batch로 사용하였다.

Table 1. Hyper parameters used for learning.
표 1. 학습에 사용된 하이퍼 파라미터

하이퍼 파라미터	지도학습	강화학습
Loss function	Mean Squared Error	Mean Squared Error
Optimizer	Adam	Adam
Learning rate	0.001	0.001
Adam_beta1	0.9	0.9
Adam_beta2	0.99	0.99
Batch_size	1024	Episode size
Epoch	100	10

학습에 소요된 시간은 기보 학습을 위한 지도학습의 경우 epoch당 대략적으로 300초 정도 소요 되었고, 모든 학습과정에는 대략 9시간이 소요되었다. 강화학습과정은 하나의 에피소드 당 평균적으로 15분이 소요되었으며, 에피소드로 모은 데이터들을 10번의 epoch로 학습시켰다. 100번의 에피소드를 마치는데 대략 25시간이 소요되었다.

2. 실험 결과

본 논문에서는 제안한 인공지능 게임 에이전트의 성능을 평가하기 위해 프로기사의 기보를 지도학습 방법으로 학습한 가치평가 네크워크인 CNN을 기반으로 한 게임 에이전트와 이를 기반으로 강화학습을 통해 학습한 게임 에이전트의 성능을 각각 실험하였다.

(1) 지도학습 네크워크 기반 에이전트 성능 평가

기보 데이터를 네크워크의 입력 상태별로 재구성한 학습데이터 7,314,328개를 자주 등장하는 데이터 5,851,472개와 자주 등장하지 않는 데이터 1,462,866개로 각각 학습데이터와 테스트데이터로 분류해 8:2의 비율로 나누어 학습을 진행하였다. 학습 진행에 따른 비용함수의 변화는 그림 6에 제시된 바와 같으며, 학습데이터에 대해 0.0638, 테스트데이터에 대해 0.0698로 수렴하였다.

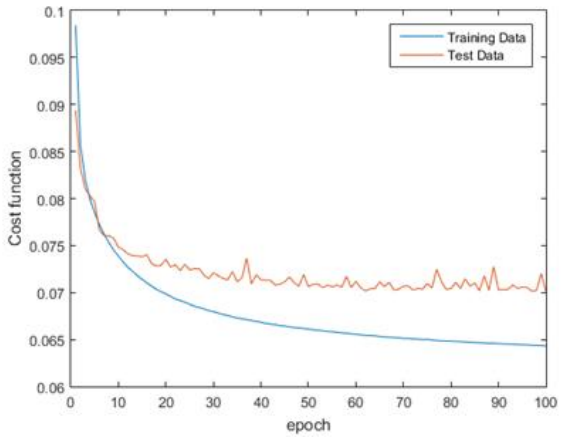

Fig. 6. Cost function in terms of learning epoch of value evaluation network.
그림 6. 가치평가 네크워크의 학습에 따른 비용함수

지도학습 네트워크의 성능 평가를 위해 학습데이터에 사용된 데이터 중, 기보의 10수에서 20수 사이의 승패통계가 50%에 근접하는 각기 다른 1000개의 상태들에 대해 이들 상태 중 하나에서 시작하여 게임이 끝날 때까지 수를 진행하는 방식으로 평가 대국을 시행하여 승패를 따지고 종합하여 최종적으로 승률을 확인하였다.

평가에 이용된 상대 에이전트는 본 연구진이 제안한 바 있는 에이전트로 비교적 간단한 형태의 네크워크 구조인 ANN 구조를 가지며, 지도학습을 통해 가치평가함수를 근사화하지만 입력으로 본 논문에서 제안하는 전략적 특성 등의 요소를 사용하지 않고 단순히 돌의 배치 상태만을 사용한 것으로 기존의 다른 인공지능 오델로 알고리즘들에 비해 우수한 성능을 보인 바 있다[1].

실험 결과 본 논문에서 제안하는 CNN 기반 게임 에이전트는 비교 대상인 ANN 기반 에이전트를 대상으로 흑일 때, 69.7%, 백일 때 72.1%의 승률을 나타내었다.

기보학습을 위한 CNN 네트워크의 성능 비교를 위해 평가대국 과정에서 등장한 상태들에 대해 사례로서 실제 기보 통계인 테이블의 가치평가 함수 $v_r(s)$와 학습된 ANN 및 CNN의 가치평가 함수 $v_\rho(s)$, $v_\sigma(s)$를 그림 7에 비교하여 제시하였다.

그림 7의 (a)는 초반부의 상태로써 백의 차례이며 기보상 승률이 50.1%인 경우이다. ANN은 이 상태가 48.9%의 승률을 가진다고 평가하였고, CNN은 49.6%의 승률을 가진다고 평가하였다. (b)는 중반부의 상태로써 흑의 차례이며 기보상 승률

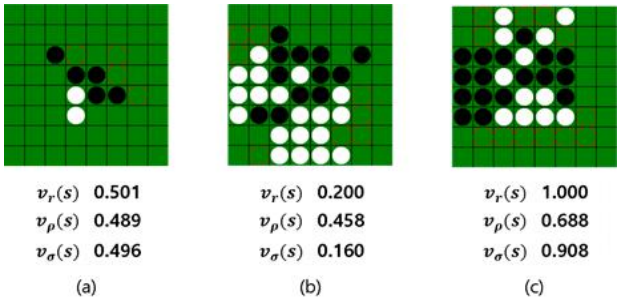

Fig. 7. Comparison of ANN and CNN based value evaluation with winning rate of real records.
그림 7. 기보 승률, ANN 및 CNN의 가치 평가 비교

이 20%로 흑이 불리한 상태인 경우이다. 이 상태에 대해 ANN 및 CNN은 각각 45.8%, 16%로 형세를 평가하여 CNN이 실제 기보에 가까운 것을 알 수 있다. (c) 역시 중반부의 상태(흑 차례)로써 실제 승률은 100%로 흑이 매우 유리한 상태이다. (c)의 상태에 대해 ANN은 68.8%로 약간 유리하다고 평가하였고, CNN은 90.8%로 제시 상황이 매우 유리한 것으로 평가하고 있다. 본 실험에서는 논문에서 제시한 예시 이외의 다른 상태들에 대해서도 전반적으로 비교 에이전트보다 우수한 결과를 얻을 수 있었다.

(2) 강화학습 네크워크 기반 에이전트 성능 평가

본 논문에서 제안하는 지도학습 네트워크를 기반으로 강화학습을 적용한 자가대국을 통해 학습한 오델로 인공지능 에이전트의 성능 평가를 위해 앞서 소개한 두 가지 지도학습 네트워크 ANN 및

Table 2. Game winning rates of the reinforcement learning agent with opponent agents in terms of learning episodes.
표 2. 강화학습 에이전트의 학습 진행에 따른 상대 에이전트와의 대국 승률

학습 에피소드	대국 상대 에이전트			
	ANN 기반		CNN 기반	
	흑	백	흑	백
10 에피소드	68%	73%	49%	51%
30 에피소드	76%	80%	55%	57%
50 에피소드	90%	91%	58%	60%
100 에피소드	100%	100%	69%	73%
150 에피소드	100%	100%	71%	74%
200 에피소드	100%	100%	76%	78%

CNN 기반의 의사결정과정을 가지는 에이전트를 평가 대상으로 하여 대국을 실시하였다. 실험 결과 강화학습 에피소드가 진행됨에 따른 자가 학습 에이전트의 비교 대상 에이전트에 대한 대국 승률을 표 9에 나타내었다.

본 실험에서는 각 에피소드 별로 100회씩 대국을 진행하였고 제안하는 인공지능 에이전트가 흑일 때와 백일 때로 구분하여 진행하였다. 총 200회의 에피소드가 진행된 후 강화학습을 적용한 자가 학습 에이전트는 ANN 기반 에이전트를 상대로 흑백 모두 100%의 승률을 보였으며, CNN 기반 에이전트를 상대로 흑백 각각 76%, 78%의 승률을 보였다.

Ⅴ. 결론

본 논문에서는 인공지능 오델로 게임 에이전트를 구현하기 위해 실제 프로기사들의 기보를 CNN으로 학습시키기 위한 구조를 제안하였고 학습된 CNN을 상태의 형세 판단을 위한 가치 평가 함수로 사용하여 최소최대탐색을 이용해 현 상태에서 최적의 수를 찾는 의사결정구조를 정책으로 사용하고 이를 강화학습을 적용하여 자가 학습하는 방법을 제안하였다.

본 논문에서 제안하는 기보 학습 방법은 기존의 ANN 기반 기보 학습 구조에 비해 네크워크의 구조로서 CNN을 사용하였으며 입력으로 단순한 돌의 배치 상태뿐만 아니라 착수 차례, 착수 가능 위치, 이동성 등 전략적 요소를 추가하였다. 본 논문에서 제안한 CNN 기반 최소최대 탐색을 적용한 게임 에이전트는 기존의 ANN을 적용한 게임에이전트와 대국을 통한 방법으로 성능을 비교하였으며, 실험 결과 흑일 때 69.7%, 백일 때 72.1%의 승률을 나타내었다.

본 논문에서는 또한 최소최대탐색과 가치평가함수를 기반으로 한 의사결정 과정을 강화학습의 정책으로 사용하여 자가 대국을 통해 학습하는 방법을 제안하였다. 실험 결과 가치 평가 네크워크 기반 강화학습이 적용된 게임 에이전트는 학습 대국이 진행될수록 강력해지는 결과를 얻을 수 있었으며, 200 회의 대국 학습 이후 ANN 기반의 게임 에이전트를 상대로 흑과 백 모두 100%의 승률을 보였고, 강화학습이 적용되기 전의 CNN 기반 지도학습만을 적용한 게임 에이전트를 상대로 흑일 때 76%, 백일 때 78%의 승률을 나타내어 성능이 개선됨을 확인할 수 있었다.

References

[1] Y. J. Jeon, Y. W. Cho, “An Implementation of Othello Game Player Using ANN based Records Learning and Minimax Search Algorithm,” *The Transactions of the Korean Institute of Electrical Engineers*, Vol.67, No.12, pp.1657-1664, 2018.
DOI: 10.5370/KIEE.2018.67.12.1657

[2] D. Silver et al., “Mastering the game of Go with deep neural networks and tree search,” *Nature* 529, pp.484-489, 2016.
DOI:

[3] D. Silver et al., “Mastering the game of Go with-out human knowledge,” *Nature* 550, pp.354-359, 2017.

[4] M. Campbell, A. J. Hoane, F. Hsu, “Deep Blue,” *Artificial Intelligence*, Vol.134, Issues 1-2, pp.57-83, 2002. DOI: 10.1016/S0004-3702(01)00129-1

[5] M. Buro, “LOGISTELLO-A Strong Learning Othello Program,” *NEC Research Institute, Princeton, NJ,* 1997.

[6] P. S. Rosenbloom, “A World-Championship-Level Othello Program,” *Artificial Intelligence*, Vol.19, Issue.3 pp.279-320, 1982.
DOI: 10.1016/0004-3702(82)90003-0

[7] K.-F. Lee, S. Mahajan, “The Development of a World Class Othello Program,” *Artificial Intelligence*, Vol.43, Issue1, pp.21-36, 1990.
DOI: 10.1016/0004-3702(90)90068-B

[8] J. Schaeffer, H. J. Herik, “*Chips Challenging Champions: Games, Computers and Artificial Intelligence*,” North Holland; 1 edition, pp.135, 2002.

[9] Gunawan et al., “Evolutionary Neural Network for Othello Game,” *Procedia-Social and Behavioral Sciences*, Vol.57, pp.419-425, 2012.
DOI: 10.1016/j.sbspro.2012.09.1206

[10] P. Liskowski, W. M. Jaskowski and K. Krawiec, “Learning to Play Othello with Deep Neural Networks,” in *IEEE Transactions on*

Games, 2018. DOI: 10.1109/TG.2018.2799997
[11] N. J. van Eck and M. van Wezel, "Reinforcement learning and its application to othello," Technical Report EI 2005-47, Econometric Institute Report, 2005.
[12] M. van der Ree and M. Wiering, "Reinforcement learning in the game of Othello: Learning against a fixed opponent and learning from self-play," *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pp.108-115, 2013.
DOI: 10.1109/ADPRL.2013.6614996
[13] R. S. Sutton, A. G. Barto, "Reinforcement Learning: An Introduction," MIT Press, Cambridge, MA, 1998.
[14] R Hahnloser, R. Sarpeshkar, M A Mahowald, R. J. Douglas, H.S. Seung, "Digital selection and analogue amplification coesist in a cortex-inspired silicon circuit," Nature. 405. pp.947-951, 2000.
[15] Y. J. Jeon, "Implementation of an artificial intelligence game agent using deep neural network and reinforcement learning," Thesis of master's degree, Seokyeong University, 2019.

BIOGRAPHY

Youngwan Cho (Member)

1991 : BS degree in Electronic Engineering, Yonsei University.
1993 : MS degree in Electronic Engineering, Yonsei University.
1999 : PhD degree in Electronic Engineering, Yonsei University.
2000~2003 : Research Engineer, Samsung Electronics.
2003~Present : Professor, Seokyeong University

Youngjin Jeon (Student Member)

2017 : BS degree in Computer Engineering, Seokyeong University.
2019 : MS degree in Electronics and Computer Engineering, Seokyeong University.