JANT 한국항행학회논문지
Journal of Advanced Navigation Technology

항행 및 항법

J. Adv. Navig. Technol. 26(2): 72-77, Apr. 2022

효율적인 경로 선택을 위한 Q-Learning 정책 및 보상 설계

Q-Learning Policy and Reward Design for Efficient Path Selection

용성중 · 박효경 · 유연휘 · 문일영*
한국기술교육대학교 컴퓨터공학과

Sung-Jung Yong· Hyo-Gyeong Park · Yeon-Hwi You· Il-Young Moon*
Department of Computer Science and Engineering, Korea University of Technology and Education, Cheonan, 31253, Korea

[요 약]

강화학습의 기법 중 Q-Learning은 주어진 상태에서 행동을 수행하면서 미래의 효율적인 기댓값을 예측하는 Q 함수를 학습하면서 최적의 정책을 학습하는 것이다. Q-Learning은 강화학습의 기본적인 알고리즘으로 많이 활용하고 있다. 본 논문에서는 Q-Learning을 바탕으로 정책과 보상을 설계하여 효율적인 경로를 선택하고 학습하는 효용성에 대하여 연구하였다. 또한 Frozen Lake 게임의 8x8 그리드 환경에 동일한 학습 횟수를 적용하여 기존 알고리즘 및 처벌 보상 정책과 제시한 처벌강화 정책의 결과를 비교하였다. 해당 비교를 통해 본 논문에서 제시한 Q-Learning의 처벌강화 정책이 통상적인 알고리즘의 적용보다 학습 속도를 상당히 높일 수 있는 것으로 분석되었다.

[Abstract]

Among the techniques of reinforcement learning, Q-Learning means learning optimal policies by learning Q functions that perform actions in a given state and predict future efficient expectations. Q-Learning is widely used as a basic algorithm for reinforcement learning. In this paper, we studied the effectiveness of selecting and learning efficient paths by designing policies and rewards based on Q-Learning. In addition, the results of the existing algorithm and punishment compensation policy and the proposed punishment reinforcement policy were compared by applying the same number of times of learning to the 8x8 grid environment of the Frozen Lake game. Through this comparison, it was analyzed that the Q-Learning punishment reinforcement policy proposed in this paper can significantly increase the learning speed compared to the application of conventional algorithms.

Key word : OpenAI Gym, Path Selection, Q-Learning, Reinforcement Learning, Reward Policy.

https://doi.org/10.12673/jant.2022.26.2.72

Received 31 March 2022; **Revised** 6 April 2022
Accepted (Publication) 18 April 2022 (30 April 2022)

***Corresponding Author ; Il-Young Moon**

Tel: +82-041-560-1493
E-mail: iymoon@koreatech.ac.kr

Ⅰ. 서 론

최근 모빌리티 시장의 규모가 증가하고 있으며, 이에 따라 차량의 완전 자동화를 위해 많은 연구가 진행되고 있다. 본 논문에서는 강화학습을 이용하여 자동주행을 위한 가장 효율적인 경로를 선택하는 방법론을 제시하고자 한다. 강화학습은 현재의 상태에 대해 최적의 행동을 선택하는 학습 방법으로, 행동에 의한 보상과 처벌을 통해 최적의 행동을 구분한다. 강화학습의 기법 중 Q-Learning은 미래의 보상 기댓값을 극대화하도록 정책을 학습하는 기법으로 Q-Learning의 기댓값을 최대로 만드는 경로가 가장 효율적인 경로가 된다. 본 논문에서는 기댓값의 정책을 처벌강화로 설정하여 기존 정책과 학습 속도에 대해 평가하고 효용성을 확인하고자 한다.

Ⅱ. 관련 기술 및 연구

2-1 Q-Learning

Q-Learning 기반 강화학습은 상태-행동 쌍에서 상태-행동 쌍으로의 전이를 고려하고, 상태-행동 쌍에 대한 가치를 학습한다. 따라서, 상태-행동 쌍에 대한 Q 값을 근사하고, 이를 기반으로 어떤 상태로부터 어떤 행동을 취할지 결정하는 방법이다. MDP(Markov Decision Process) 이론을 기반으로 하는 대표적인 알고리즘[1]이다. 탐험하기 위해 입실론 탐욕적 정책(ε-Greedy Policy)을 주로 이용한다. 탐욕적 정책이란 특정 상태에서의 확률로 랜덤한 행동을 선택하고, $(1-\varepsilon)$의 확률로 제일 높은 Q 값을 가지는 행동을 선택하는 방법이다. 해당 알고리즘은 수식(1)과 같이 정의된다[2].

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_{\alpha} Q(S_{t+1}, a) - Q(S_t, A_t)\right] \quad (1)$$

각 시간 t에서 Agent는 상태(S_t)에서 행동(A_t)을 수행하고, 새로운 상태(S_{t+1})로 전이되고, 이를 통해 보상(R_t)을 획득한다. 이 과정을 이전의 값과 새 정보의 가중 합을 이용해 반복하는 것을 의미하고 있다.

학습된 행동 가치 함수 Q는 자신이 따르는 정책에 상관없이 최적 행동 가치 함수 q*를 직접적으로 근사한다.

2-2 DQN (Deep Q-Network)

DQN은 Q-Learning과 딥러닝을 합친 것을 말한다. 2013년, 영국의 스타트업 Deep Mind에서 선보인 기술이다. 심층 강화학습 기술을 이용해 다양한 Atari 게임에서 사람보다 더 게임을 잘하도록 한 것이며, 현재도 이와 관련한 연구가 활발하게 이루어지고 있다. Atari 게임이란 미국 비디오 게임 개발사 Atari에서 개발한 비디오 아케이드 게임을 의미한다.

이 DQN은 큐 테이블(Q-table) 대신 신경망(Neural Network)을 사용해서, 뉴럴넷이 Q 가치를 근사해낼 수 있도록 학습시키는 것이다. 이러한 이유로 DQN은 근사기(Approximator) 또는 근사 함수(Approximating Functioin)이라고 부르기도 한다. 이미지 인식을 위해 합성곱 신경망을 이용하고, 샘플 간 상관관계를 없애며 샘플 효율성을 높이기 위해 경험 리플레이를 도입했다는 특징이 있다. 또한, 학습의 안정성을 위해 Agent의 행동을 결정하는 Online Q-Network와 목표 Q 값 계산에 사용되는 target Q-network를 분리하여 사용하였다는 특징이 있다[3].

Ⅲ. 연구방법

3-1 연구환경

본 논문은 강화학습 라이브러리 OpenAI Gym을 이용하여 Q-Learning 보상에 대한 설계를 제안하고 시뮬레이션을 진행하였다.

OpenAI Gym이란 OpenAI에서 만든 라이브러리 패키지로, 이를 통해 강화학습 알고리즘(Reinforcement Learning Algorithm)을 개발하고, 훈련을 수행할 수 있는 Agent와 환경을 제공 받을 수 있다[4]. OpenAI Gym의 Frozen Lake 환경은 출발 지점부터 도착지점까지의 함정에 빠지지 않고 이동하여 도착지점까지 이동하는 학습을 하는 시뮬레이션이다. 효율적인 경로 선택 강화학습의 시뮬레이션으로 많은 연구자가 사용하는 환경으로 본 논문에서도 제안된 알고리즘을 평가하기 위해 선택하게 되었다.

Frozen Lake 환경은 에이전트(Agent)가 무작위 행동(Action)을 하고 행동에 따라 환경(Environment) 속에서 상태가 변경되고 상태(State)에 따라 보상이 주어지는 환경을 제공한다. 에이전트는 시작점(S)에서 출발하여 함정(H)에 빠지게 되면 게임은 종료가 되고, 함정(H)를 제외한 얼은 면(F)를 지나 목표 지점(G)에 도착하는 게임으로 목표 지점에 도착하면 보상이 주어지게 되고 보상 값에 따라 효율적 경로를 선택하는 강화학습 환경이다.

3-2 Q-Learning 처벌강화 정책 제안

1) 처벌강화 방법

기존 Frozen Lake의 보상방식은 수식(2) 큐 함수에 대한 벨만 최적 방정식(The Bellman Equation)의 알고리즘에 따라 에이전트가 목표지점에 도달했을 때 보상 값을 지급하고 경로 선택 시 최고의 보상 값을 선택하도록 하였다[5].

$$q^*(s,a) = E\left[R_{t+1} + \Upsilon \max_{a'} q^*(S_{t+1}, a')\ S_t = s, A_t = a\right] \quad (2)$$

표 1. Frozen Lake 게임 학습 대상
Table 1. Frozen Lake game learning target

Category	Contents
State	Agent position coordinates (x,y)
Action	One of the actions, top, bottom, left and right
Rewards or Penalty	Previous Q value in that direction when the agent target point is reached = 1, All surrounding Q values when the agent falls into a trap = −1

본 논문에서는 수식(3)과 같이 효율적인 경로 선택을 위한 처벌강화 정책을 제안하였다. 표1과 같이 처벌강화 정책은 에이전트가 함정(H)에 빠지게 되면, 함정 주변의 얼은 면(F)의 가치함수(보상) 처벌 값을 적용하여 함정을 회피하여 목표 지점까지 빠르게 도달하는 학습 성공률을 높이고자 한다.

$$Q(s,a+x) = -1, Q(s,a-x) = -1 \quad (3)$$

상태 s에서 행동 a를 진행하였을 때, 가치함수 Q(s,a) 사용한다. 이 때, 미래의 보상 및 처벌을 위해서 Q 값으로 −1, 0, 1의 값을 사용한다. Q 값은 Agent의 Action으로 얻은 보상과 처벌 값으로 갱신한다.

Agent가 함정(H)에 빠질 경우, 주변 모든 얼은 면(F)의 해당 방향의 Q 값을 −1로 갱신하고, Agent가 목표 지점(G)에 도달할 경우, 해당 방향의 이전 Q 값을 1로 갱신한다.

최종적으로 Agent가 행동 a를 진행할 때, Q 값의 새로운 정보를 이용하여 경로를 선택한다.

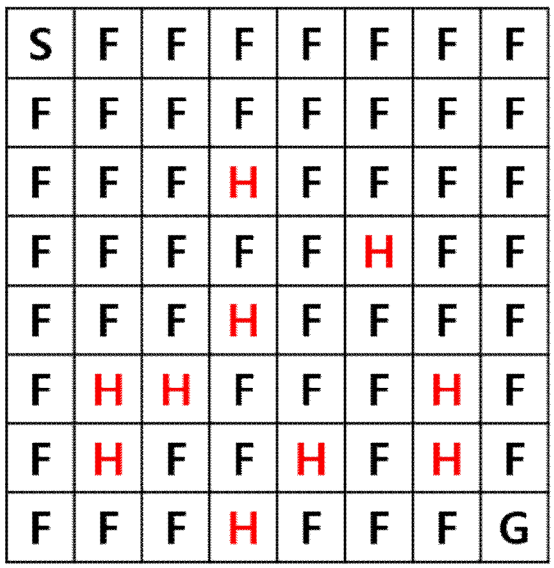

그림 1. Frozen Lake 시뮬레이션 환경
Fig. 1. Frozen Lake simulation environment

2) 시뮬레이션 환경 구성

시뮬레이션 환경은 Frozen Lake의 일반적인 환경 4x4 그리드 환경에서는 적은 학습 횟수를 진행하였을 때 성공률에 대한 확률값이 크지 않으므로 본 논문에서는 8x8 그리드 환경과 10개의 함정을 적용하였다. 그림 1과 같이 각 그리드는 시작점(S), 얼은 면(F), 함정(H), 목표지점(G)로 배치하고, Agent가 시작점(S)로부터 연결된 얼은 면(F)을 지나 목표지점(G)까지 도달하는 경로를 학습하게 된다. 또한 제시한 알고리즘의 성공률 비교를 위해 수식(2)의 Q 함수에 대한 벨만 최적 방정식 알고리즘, 처벌 보상 정책의 시뮬레이션 환경도 구성하여 진행하였다.

Ⅳ. 연구결과

4-1 Q-Learning 보상 정책 시뮬레이션 결과

기존 Q-Learning 학습환경은 4x4 그리드 환경에서 기댓값 보상 정책을 진행하는 알고리즘을 적용하였지만 본 논문에서는 제시한 처벌강화 정책의 학습 속도와 동일한 환경에서 비교하기 위해 8x8 그리드 환경에서 학습 시뮬레이션을 진행하였다. 그림 2와 같이 Frozen Lake 환경에 100번 학습을 진행하였고, 그림 3과 같이 결과를 확인할 수 있었다. 짧은 학습 횟수이기 때문에 기존 방법으로는 에이전트가 목표 지점까지 도달할 수 없었다. 학습 횟수가 늘어난다면 성공률을 높아지게 된다.

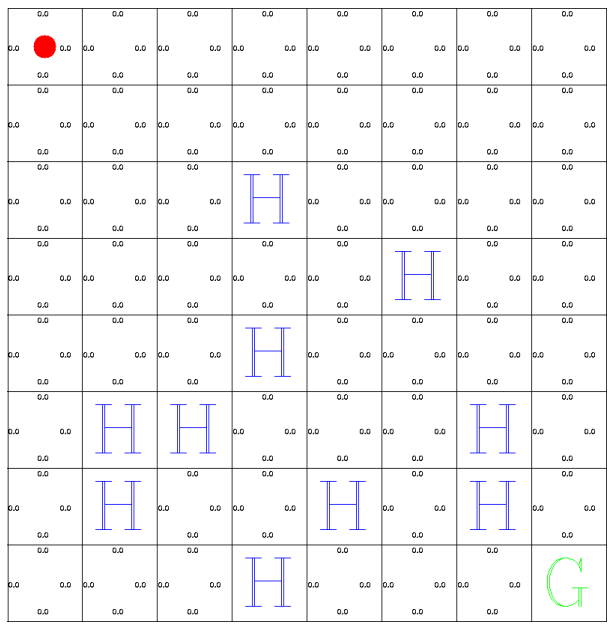

그림 2. Q-Learning 보상 정책 시뮬레이션 결과
Fig. 2. Q-Learning Compensation Policy Simulation Result

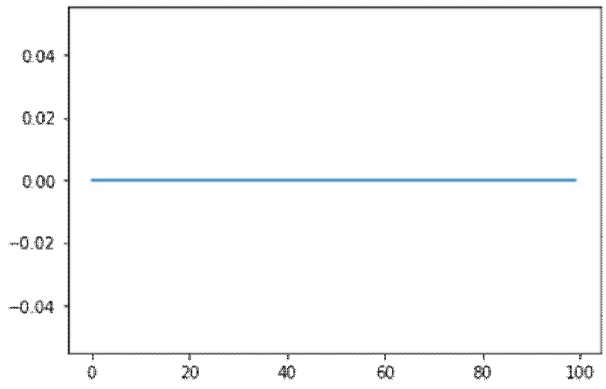

그림 3. Q-Learning 보상정책 시뮬레이션 학습 성공률
Fig. 3. Q-Learning Compensation Policy Simulation Learning Success Rate

4-2 장애물 보상 Q-Learning 시뮬레이션 결과

에이전트가 목표물 도달 시 성공 및 장애물 보상 정책을 진행하는 알고리즘을 적용하여 기존 Q-Learning 보상 정책 시뮬레이션과 동일한 환경으로 학습을 진행하였다. 그림 4와 같이 8x8 그리드 환경과 100번 학습을 진행하여 그림 5와 같이 결과를 확인할 수 있었다. 장애물 보상 시뮬레이션 결과 50%의 성공률을 보여 Q-Learning 보상 정책 알고리즘보다 학습 속도가 높아진 것을 확인할 수 있었다.

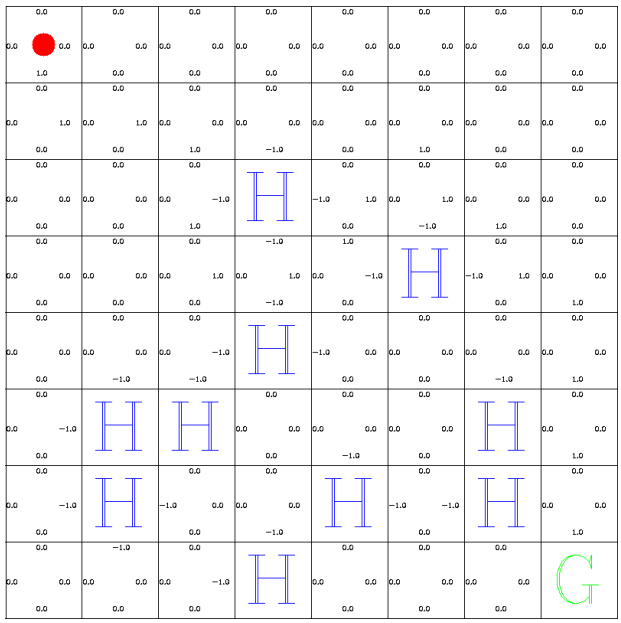

그림 4. Q-Learning 장애물 보상 시뮬레이션 결과
Fig. 4. Q-Learning Obstacle Compensation Simulation Result

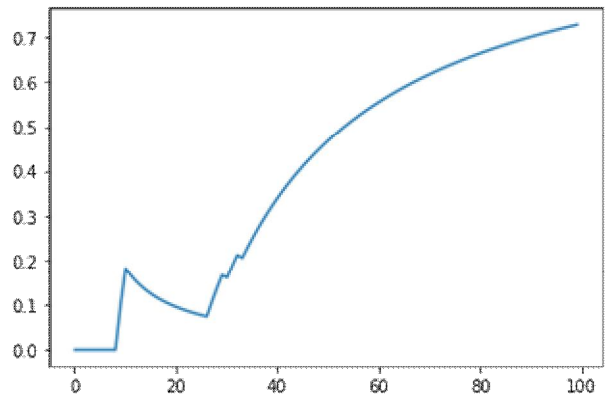

그림 5. Q-Learning 장애물 보상 시뮬레이션 학습 성공률
Fig. 5. Q-Learning Obstacle Compensation Simulation Learning Success Rate

4-3 처벌강화 Q-Learning 시뮬레이션 결과

본 논문에서 제안한 Q-Learning 학습은 장애물 보상에 대해 장애물 인접 방향의 처벌 보상 값을 모두 할당하여 에이전트가 함정에 한 번 빠지고 주변의 얼은 면에서 함정 방향의 Q값을 처벌보상 값으로 대입하여 에이전트가 함정을 회피하도록 한다. 시뮬레이션 환경은 그림 6과 같이 이전의 환경과 같이 8x8 그리드 환경과 함정의 위치를 동일하게 100번의 학습 시뮬레이션을 진행하였다. 그림 7과 같이 학습 성공률은 85%로 이전 알고리즘보다 학습 속도가 빨라진 것을 확인할 수 있었다.

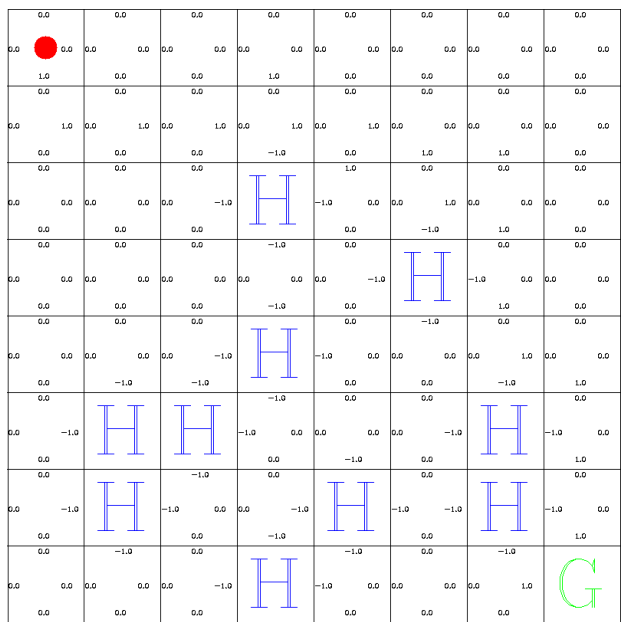

그림 6. 처벌강화 Q-Learning 시뮬레이션 결과
Fig. 6. Strengthening Punishment Q-Learning Simulation Result

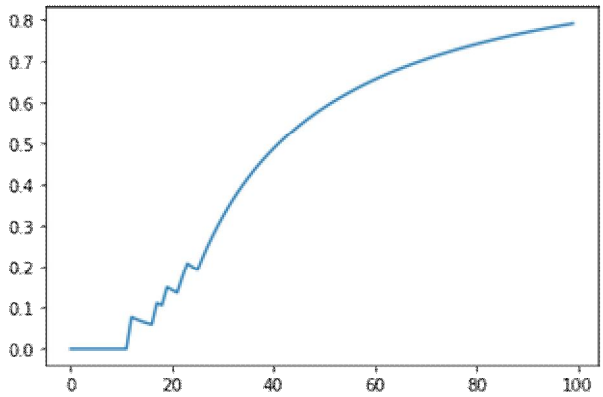

그림 7. 처벌강화 Q-Learning 시뮬레이션 학습 성공률
Fig. 7. Strengthening punishment Q-Learning Simulation Learning Success Rate

Ⅴ. 결 론

강화학습은 현재의 상태에 대해 최적의 행동을 선택하는 학습 방법으로, 행동에 의한 보상과 처벌을 통해 최적의 행동을 구분한다. Q-Learning 기반 강화학습은 상태-행동 쌍에서 상태-행동 쌍으로의 전이를 고려하고, 상태-행동 쌍에 대한 가치를 학습한다. 따라서, 상태-행동 쌍에 대한 Q 값을 근사하고, 이를 기반으로 어떤 상태로부터 어떤 행동을 취할지 결정하는 방법이다. 또한, Q-Learning은 미래의 보상 기댓값을 극대화하도록 정책을 학습하는 기법으로 Q-Learning의 기댓값을 최대로 만드는 경로가 가장 효율적인 경로가 된다. 본 논문에서는 기댓값의 정책을 처벌강화로 설정하여 기존 알고리즘의 보상 정책과 학습 속도에 대해 시뮬레이션을 통해 결과를 확인할 수 있었다. 본 논문에서 제안한 처벌강화 정책을 통해 적은 학습 횟수에도 불구하고 기존 보상 및 처벌 보상 정책에 비해 높은 학습 성공률을 보여주었다. 한정된 시뮬레이션 환경으로 인해 차후 연구에서는 다양한 환경을 구성하여 시뮬레이션 평가 또는 시뮬레이션 평가도구를 개발할 수 있도록 하고, 장애물의 난이도에 따른 보상, 처벌정책에 따라 장애물을 통과하거나 회피할 수 있는 방법을 연구하여 효율적 경로를 선택할 수 있는 강화학습을 진행할 것이다.

Acknowledgments

본 연구는 2021년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력 기반 지역혁신 사업의 결과입니다. (2021RIS-004)

References

[1] Watkins, C.J.C.H., Dayan, P., “Q-learning”, *Machine Learning*, Vol. 8, No. 1, pp. 279-292, May. 1992.
[2] Watkins, C.J.C.H, Learning from Delayed Rewards, Ph.D. thesis, King’s College, London, May. 1989.
[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing Atari with Deep Reinforcement Learning”, arXiv preprint arXiv, 1312.5602, Dec. 2013.
[4] G. Brockman, V.Cheung, L. Pettersson, J. Schneider, J.Schulman, J.Tang, and W. Zaremba, “OpenAI Gym”, arXiv preprint arXiv, 1606.1540, Jun. 2016.
[5] Clifton, J., and Laber, E., “Q-Learning: Theory and Applications”, *Annual Review of Statistics and Its Application*, Vol. 7, No. 1, pp. 279-301, Mar. 2020.

용 성 중 (Sung-Jung Yong)
2020년 8월 : 한국기술교육대학교 대학원 컴퓨터공학과 공학석사
2021년 8월 ~ 현재 : 한국기술교육대학교 대학원 컴퓨터공학과 박사과정
※관심분야 : AI, 빅데이터, 추천 시스템, 웹 등

박 효 경 (Hyo-Gyeong Park)
2021년 8월 :한국기술교육대학교 컴퓨터공학 학사
2021년 8월 ~ 현재 : 한국기술교육대학교 대학원 컴퓨터공학과 석사과정
※관심분야 : AI, 빅데이터, 추천 시스템 등

유 연 휘 (Yeon-Hwi You)
2016년 3월 ~ 현재 :한국기술교육대학교 컴퓨터공학 학사과정
※관심분야 : AI, 빅데이터, 추천 시스템 등

문 일 영 (Il-Young Moon)
2005년 2월: 한국항공대학교 항공통신정보공학과 공학박사
2005년 3월 ~ 현재: 한국기술교육대학교 컴퓨터공학과 교수
※관심분야 : AI, 무선인터넷 응용, 무선 인터넷, 모바일IP 등