에너지 하베스팅 다중사용자 네트워크를 위한 강화학습 기반 실시간 전력 최적화 기법

김혜성, 신원재
아주대학교
heasungkim@ajou.ac.kr, wjshin@ajou.ac.kr

Reinforcement Learning-Based Real-Time Power Allocation For Energy Harvesting Multiuser Networks

Heasung Kim, Wonjae Shin
Ajou Univ.

요 약

최근 재생 가능한 에너지에 관한 관심의 증대 하에, 그러한 에너지 자원으로부터 전력을 공급받아 네트워크를 이루는 에너지 하베스팅 네트워크가 저전력 네트워크의 후보로 각광을 받고 있다. 그러나 안정적인 전력 공급이 보장되지 않는 상황에서 일반적인 전력 할당 정책을 사용하면 네트워크의 정보 전송 효율이 저하될 수 있다. 이러한 문제점을 극복하기 위하여, 본 논문에서는 경사 하강법 기반 강화학습을 사용하여 시스템의 실시간 상태에 기반한 전력 할당 정책을 학습하는 방식을 제안한다. 특히 경사 하강법과 함께 사용되는 근사 함수의 설계가 강화학습 알고리즘의 성능에 유의미한 영향을 미칠 수 있음을 실험을 통하여 보인다.

Ⅰ. 서 론

네트워크의 전 지구적 보급을 위해 각광받는 기술 중 하나인 에너지 하베스팅 통신은, 전력 공급망을 통한 안정적인 공급보다 태양열, 풍력 등의 수확 가능한 에너지를 주 에너지원으로 사용하는 통신시스템이다. 해당 통신 시스템에서는 일정 수준 이상의 전력이 지속적으로 공급된다는 보장을 할 수 없다는 것이 특징으로, 통신을 위한 전력의 효율적인 사용이 필수적이다.

제한된 전력으로 정보 전송량을 최대화 하기 위해서, 본 논문에서는 하나의 에너지 하베스팅 송신기와 다중 사용자가 가정된 시스템 모델(그림 1.)을 위한 전력 할당 정책을 제안한다. 에너지 하베스팅 송신기는 에너지 수확 및 유한한 에너지 보유량의 조건 하에, 다중 사용자의 정보 전송량 최대화를 목표로 전력 할당 정책을 학습한다. 전력 할당 정책 학습을 위해서 강화학습(Reinforcement Learning) 알고리즘 중 하나인 Deep Deterministic Policy Gradient(DDPG)[1]가 적용되었으며, 심층 인공신경망[2]과 격자 네트워크[3]로 학습된 전력 할당 정책의 성능을 다양한 실험을 통해 나타낸다. 특히 전력 할당 정책을 위해 크기가 제한된 격자 네트워크는 임의로 설계된 심층 인공신경망보다 높은 성능을 달성하는 정책을 안정적으로 학습해 낼 수 있음을 보인다.

Ⅱ. 본론

가. 시스템 모델

전력 $b_{\max}$를 저장할 수 있는 배터리가 탑재된 에너지 하베스팅 송신기는 이산적 매 타임 슬롯마다 k번째 사용자에게 신호를 전송하기 위해 사용할 전력량 $\mathbf{p}_k \in \mathbb{C}^{N\times 1}$를 정한다. Multiple-input Single-output(MISO) 채널을 가정할 경우 각 사용자를 위한 채널은 $\mathbf{h}_k \in \mathbb{C}^{N\times 1}$ 로 표현될 수 있다. k $(k=1,...,K)$ 번째 사용자가 달성할 수 있는 최대 정보 전송량은 아래와 같이 표현된다.

$$R_k = \log_2\left(1+\frac{|\mathbf{h}_k^H \mathbf{p}_k|^2}{\sum_{j=1, j\neq k}^{K} |\mathbf{h}_k^H \mathbf{p}_j|^2 + \sigma^2}\right) \quad (1)$$

이 때 σ^2은 표준 정규분포를 따르는 부가 백색 가우스 잡음(additive white Gaussian noise)의 분산을 나타낸다. 가독성을 위해 전송시간 T를 1로 간단히 정하면, 송신기의 시간 i일 때 에너지 보유량 b_i는 전력 $p_i(\leq b_i)$를 사용하는 경우 다음 시간 슬롯에서의 에너지 보유량은 아래와 같이 계산된다.

$$b_{i+1} = \min(b_{\max}, b_i - p_i + e_i) \quad (2)$$

이 때 e_i는 시간 i일 때 에너지 수확량을 나타낸다. 즉 최대 에너지 보유 가능량을 넘지 않는 선에서 에너지를 저장할 수 있고, 매 시간 슬롯마다 다중 사용자에게 총 전력 p_i을 사용하여 신호를 전송한다.

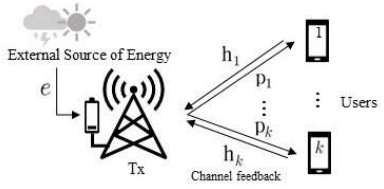

그림 1. 에너지 하베스팅 송신기와 다중 사용자

이 때 시스템 모델은 무한한 동작 시간 동안 감가된 정보 전송량 합을 최대화 하는 목적 함수를 가지며, 전력 제한 (2)로 인한 제약으로 인하여 아래와 같은 최적화 문제를 형성한다.

$$(\mathrm{P1}) \quad \max_{\mathbf{p}_k} E\left[\sum_{i=0}^{\infty}\gamma^i\left(\sum_{k=1}^{K} R_k\right)\right]$$
$$\text{s.t.} \sum_{k=1}^{K}\|\mathbf{p}_k\|^2 \leq p_i, \quad (2)$$

위와 같은 문제를 Markov Decision Process로 해석하여 연속적 전력 할당 정책을 학습하는 DDPG 알고리즘[1]을 이용한다.

나. DDPG와 근사함수를 이용한 강화학습 접근법

DDPG를 이용하기 위하여 $N=1$인 경우를 가정하여 상태를 $s=(e,\|\mathbf{h}_1^H\|,...,\|\mathbf{h}_K^H\|,b)$로 정의하였다. 상태를 입력으로 하여 해당 시간 슬롯의 총 전력량 p를 결정하는 정책 함수 $\pi_{\theta^\mu}(s)$를 정의하고 이를 이루는 파라미터를 학습한다. 이 때 근사함수를 이루는 파라미터의 집합 θ^μ의 값들은 경사 상승 방향 $E[\nabla_{\theta^\mu}Q(s,a;\theta^Q)]$으로 업데이트 된다[1]. 이 때 $Q(s,a;\theta^Q)$는 근사된 행동-가치 함수를 의미하며, θ^Q는 행동-가치 함수를 이루는 모든 파라미터의 집합이다. 집합 θ^Q의 모든 파라미터들은 행동-가치함수의 평균제곱오차를 최소화하는 방향으로 경사 하강법을 통하여 업데이트된다. 전력 할당 정책 $\pi_{\theta^\mu}(s)$는 지정된 시간 슬롯에서 사용할 총 전력량 p_i를 결정한다. 결정된 총 전력량은 총 K명의 사용자에게 동일하게 분배된다. 학습 알고리즘은 [1]에서 명시된 학습방법을 차용하며, 10회의 학습동안 정책 성능 향상이 관찰되지 않을 경우 학습 알고리즘을 종료한다.

본 논문에서는 근사함수의 형태에 따른 알고리즘 안정성을 실험하기 위하여 두가지 근사함수를 사용한다. 첫째로 실험된 심층 인공신경망은 256-128의 hidden-layer로 구성되었고, 모든 파라미터는 0.0의 평균과 16의 표준편차를 가지는 가우시안 분포를 따른다[2]. 모든 레이어는 sigmoid 함수의 activation을 가진다. 둘째로 실험된 격자 네트워크는 8개 의 균등한 부분으로 $s=(e,\|\mathbf{h}_1^H\|,...,\|\mathbf{h}_K^H\|,b)$의 모든 원소 구간을 나누어 단위 크기로 선형 변환을 하는 교정(calibration) 층과, 2의 크기를 가지는 다중선형적 보간(interpolation) 층 Φ으로 이루어져있다. $t(\leq 2^{2+K})$번째 보간층은 아래와 같이 나타낼 수 있으며[3], $v_t[d]\in\{0,1\}^{2+K}$이다.

$$\Phi_t(s)=\prod_{d=1}^{2+K}s[d]^{v_t[d]}(1-s[d])^{1-v_t[d]},\qquad t=1,...,2^{2+K} \quad (3)$$

정책을 이루는 격자 네트워크의 최대와 최솟값은 각각 0과 $b_{\max}$로 제한되었고, 보간 층은 학습 가능한 보간 파라미터와의 곱을 통하여 정책을 이룬다. 심층 인공신경망과 격자 네트워크는 동일한 구조로 정책과 행동-가치 함수를 위해 설계되었다.

다. 시뮬레이션 결과

채널 $\mathbf{h}_k$ 은 분산 1.0을 가지는 complex Gaussian distribution를 따르며 $K=2$로 가정하였다. 에너지 하베스팅 송신기의 최대 전력 저장량은 2로 가정하였고, 에너지는 p^h의 확률로 최대 전력 저장량만큼 수확이 가능한 베르누이 분포를 가정하였다.

그림 2.는 에너지 수확 확률에 따른 격자 네트워크와 DDPG를 사용한 강화학습 기반 전력 할당 정책이 달성 가능한 정보 전송량을 나타낸다. Greedy Policy의 경우 사용 가능한 에너지를 매 타임 슬롯마다 모두 소모하는 정책을 의미한다. 시스템의 상태를 고려한 전력 운영을 통해서 단순한 전력 할당 정책으로 달성할 수 있는 정보 전송량을 최대 10.9% 이상 ($p^h=0.5$) 증가시킬 수 있음이 관찰된다.

그림 3.은 에너지 수확 확률이 0.5인 통신 환경에서 심층 인공신경망과 격자 네트워크가 달성한 성능의 분포를 나타낸다. 각 실험은 완전히 독립적인 환경에서 500회씩 이루어졌으며, 각 회당 Ⅱ-나.에서 언급된 학습방법을 통하여 최종 성능이 결정되었다. 심층 인공신경망과 격자 네트워크는

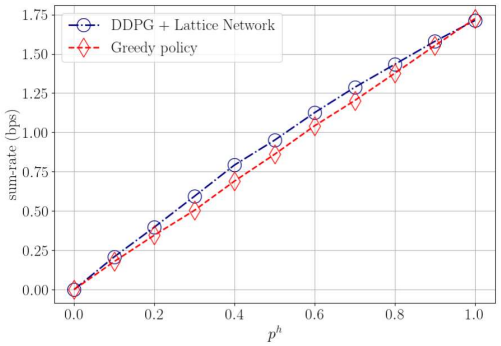

그림 2. 에너지 수집 확률에 따른 DDPG와 격자 네트워크를 기반으로 학습된 전력 할당 정책의 성능.

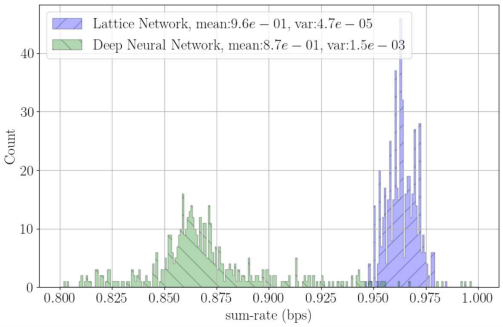

그림 3. 심층 인공신경망과 격자 네트워크를 이용하여 학습한 정책의 정보 전송량 분포.

동일한 강화학습 알고리즘과 함께 사용되었음에도 불구하고, 임의의 통신 환경에서 격자 네트워크를 통한 정책 성능의 분산이 심층 인공신경망이 달성한 성능의 분산보다 작은 값을 가지는 것을 확인하였다.

Ⅲ. 결론

본 논문은 에너지 하베스팅 통신 시스템에서 다중 사용자의 정보 전송량 최대화를 위한 강화학습 접근법을 다루었다. DDPG 알고리즘은 심층 인공신경망과 격자 네트워크와 결합되어 실험이 이루어졌다. 정책 학습을 위해 크기가 제한된 격자 네트워크는, 심층 인공신경망이 평균적으로 달성 가능한 정보 전송량보다 높은 정보 전송량을 달성하였다. 이는 근사 함수를 기반으로 한 강화학습 접근법의 설계 시 시스템 모델의 성능과 정책의 구조적 특징을 고려하는 것이 전력 할당 정책의 성능과 안정성을 크게 향상시킬 수 있음을 암시한다.

ACKNOWLEDGMENT

This work is supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-00467, Intelligent 6G Wireless Access System).

참 고 문 헌

[1] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[3] M. Gupta, A. Cotter, J. Pfeifer, K. Voevodski, K. Canini, A. Mangylov,W. Moczydlowski, and A. Van Esbroeck, "Monotonic calibrated interpolated look-up tables," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3790 - 3836, 2016.