Regular paper | Journal of the KIECS. pp. 1153-1160, vol. 16, no. 6, Dec. 31. 2021, t. 110, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2021.16.6.1153

SDN에서 심층강화학습 기반 라우팅 알고리즘

이성근*†

A Routing Algorithm based on Deep Reinforcement Learning in SDN

Sung-Keun Lee*†

요 약

본 논문은 소프트웨어 정의 네트워크에서 심층강화학습을 활용하여 최적의 경로를 결정하는 라우팅 알고리즘을 제안한다. 학습을 위한 심층강화학습 모델은 DQN 을 기반으로 하고, 입력은 현재 네트워크 상태, 발신지, 목적지 노드이고, 출력은 발신지에서 목적지까지의 경로 리스트를 반환한다. 라우팅 작업을 이산 제어 문제로 정의하며, 라우팅을 위한 서비스 품질 파라미터는 지연, 대역폭, 손실률을 고려하였다. 라우팅 에이전트는 사용자의 서비스 품질 프로파일에 따라 적절한 서비스 등급으로 분류하고, SDN에서 수집된 현재 네트워크 상태로부터 각 링크 별로 제공할 수 있는 서비스 등급을 변환한다. 이러한 변환된 정보를 토대로 발신지에서부터 목적지까지 요구되는 서비스 등급을 만족시키는 경로를 선택하도록 학습을 한다. 시뮬레이션 결과는 제안한 알고리즘이 일정한 에피소드를 진행하게 되면 올바른 경로를 선택하게 되고, 학습이 성공적으로 수행됨을 나타냈다.

ABSTRACT

This paper proposes a routing algorithm that determines the optimal path using deep reinforcement learning in software-defined networks. The deep reinforcement learning model for learning is based on DQN, the inputs are the current network state, source, and destination nodes, and the output returns a list of routes from source to destination. The routing task is defined as a discrete control problem, and the quality of service parameters for routing consider delay, bandwidth, and loss rate. The routing agent classifies the appropriate service class according to the user's quality of service profile, and converts the service class that can be provided for each link from the current network state collected from the SDN. Based on this converted information, it learns to select a route that satisfies the required service level from the source to the destination. The simulation results indicated that if the proposed algorithm proceeds with a certain episode, the correct path is selected and the learning is successfully performed.

키워드

QoS Aware Routing Algorithm, Software Defined Network, Deep Reinforcement Learning, QoS Profile
서비스 품질 기반 라우팅 알고리즘, 소프트웨어 정의 네트워크, 심층강화학습, 서비스 품질 프로파일

Ⅰ. 서 론

소프트웨어 정의 네트워크(SDN :Software Defined Network)는 새로운 네트워킹 패러다임으로 요구에 따른 자원 할당, 손쉬운 재구성 및 프로그래밍 가능한 네트워크 관리 기능 등을 제공한다[1]. SDN은 네트워크 상태에 대한 중앙 집중화된 정보를 제공함으로써 네트워크 관리와 제어를 더욱 유연하고 일관성 있고,

* 순천대학교 멀티미디어공학과(sklee@scnu.ac.kr)
• 접 수 일 : 2021. 10. 22
• 수정완료일 : 2021. 11. 19
• 게재확정일 : 2021. 12. 17
• Received : Oct. 22, 2021, Revised : Nov. 19, 2021, Accepted : Dec. 17, 2021
• Corresponding Author : Sung-Keun Lee
Dept. Multimedia Eng., Sunchon National University,
Email : sklee@scnu.ac.kr

총괄적으로 수행할 수 있으며, 사용자 연결 서비스를 위해 할당된 대역폭과 경로를 동적으로 조정함으로써 트래픽 제어와 관리 기능을 향상시킬수 있다[2]. SDN에서 네트워크 관리 기능은 논리적으로 제어 평면과 데이터 평면으로 구성된다. SDN 제어 평면의 라우팅 모듈은 실시간으로 네트워크 상태 정보를 수집한다. 상태 정보를 기반으로 사용자가 요구하는 지연, 손실 및 대역폭 등의 서비스 품질(Quality of Service : QoS) 파라미터 정보에 따라 통신을 수행하는 양단간에 QoS 를 만족하는 경로를 제공한다. 다양한 네트워크 상태를 고려한 QoS 라우팅에 대한 많은 연구가 수행되었고, 이러한 연구의 대부분은 모델 기반이며, 사용자 요구와 네트워크 환경을 적절히 모델링할 수 있다는 가정하에 연구가 수행되었다[3][4]. 또한 여러 QoS 매개 변수를 처리하려면 높은 수준의 컴퓨팅 자원이 필요하다. 현재의 통신 네트워크는 동적인 특성으로 지니고 있으며, 매우 복잡하게 진화됨에 따라 모델링과 제어가 어렵다.

DeepMind에서 DQN(Deep Q-Network)을 제안한 이후, 심층강화학습(Deep Reinforcement Learning : DRL) 방법은 경험을 통해 학습하기 때문에 정확한 수학 모델링 과정이 필요 없고, 매우 복잡한 문제를 해결할 수 있기 때문에 다양한 분야에 적용되고 있다. 전통적 강화학습은 복잡한 상태와 행동 집합이 필요한 대규모 시스템에 적용하기에는 한계점을 나타낸다. 심층강화학습은 강화학습 이론에 딥러닝을 결합함으로써 전통적 강화학습이 직면한 한계를 극복할 수 있다. 심층강화학습은 통신 및 네트워킹 분야에서 제기되는 대규모의 복잡한 문제를 해결할 수 있는 능력이 있기 때문에 다양한 분야의 연구자들이 관심을 가지게 되었다[6]. SDN 환경에서 강화학습을 적용하여 QoS 인식 적응형 라우팅 알고리즘을 제안하였다[7]. 최근의 몇몇 연구는 DDPG(Deep Deterministic Policy Gradient) 등 다양한 DRL 알고리즘을 통신 네트워크의 라우팅에 적용하였으며, 라우팅 문제를 연속 제어 문제로 고려하였다[8]. 즉, 지속적인 트래픽 흐름 트래픽 매트릭스를 사용하여 최단 경로를 결정한다. 이러한 DRL 방법은 각 발신지-목적지 쌍 간의 통신을 위해 k-최단 경로만 고려하였다. 따라서 더 나은 서비스 품질을 제공할 수 있는 다른 경로가 존재할 수 있기 때문에 성능이 제한될 가능성이 존재한다.

본 논문에서는 라우팅 문제를 이산 제어 문제로 정의한다. 통신을 원하는 송신자는 목적지 정보 뿐만 아니라 원활한 서비스를 제공 받기 위해 필요한 최소한의 서비스 품질 파라미터를 서비스 제공자에게 전달한다. 서비스 품질 파라미터는 매핑 테이블 또는 매핑 함수를 통하여 적절한 서비스 등급으로 변환된다. SDN 제어기는 현재의 네트워크 상태 정보를 각 링크별로 수집하고, 해당 링크의 상태 정보를 서비스 등급으로 변환한 정보를 가지고 있다. DRL 에이전트는 서비스 등급 정보를 토대로 발신지-목적지 쌍 간의 QoS 요구 조건을 만족하는 통신을 위한 경로를 찾을 수 있도록 학습한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 분석하고, 3장에서는 시스템 모델과 문제 정의에 대해 설명한다. 4장에서는 제안한 DQN 기반의 라우팅 알고리즘에 대해 설명하고, 현재까지 진행된 성능분석 결과를 나타내고, 5장에서는 본 논문의 결론을 맺는다.

II. 관련 연구

2.1 심층강화학습

강화학습은 순차적 행동 결정 문제를 풀기 위해 최적의 정책을 구하는 과정이다. 강화학습은 에이전트와 환경이라는 두 개의 개체로 구성되며, 이들 간의 상호작용은 지속적으로 환경에 영향을 미치고, 에이전트는 환경과의 상호 작용을 통해 얻게 되는 보상값을 통해 학습한다[5]. 그림 1에 나타낸 바와 같이 에피소드 동안에, 에이전트는 상태 정보 s_t 를 관찰하고, 각 상태에서 정의된 정책 $\pi(s_t)$ 에 따라 행동 a_t 를 결정한다. 에이전트는 행동을 수행하고, 환경으로부터 스칼라 형태의 보상값 $R(s_t,a_t)$을 받고, 환경의 변화된 다음 상태를 관찰한다.

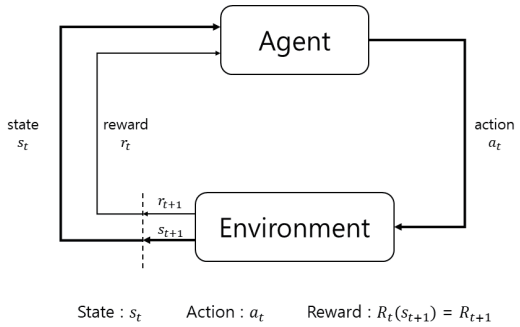

그림 1. 심층강화학습 구성요소
Fig. 1 Components of deep reinforcement learning

에이전트는 전체 에피소드 동안 환경으로부터 얻어지는 누적 보상값을 최대화하는 정책을 찾도록 학습한다. 정책 $\pi(s_t)$ 는 특정 상태에서 취할 수 있는 모든 행동들에 대한 확률값을 나타내며, 출력 행동을 이산 행동 공간 P(a)에 대한 확률 분포로 모델링 할 수 있다. 따라서 에이전트가 특정 상태에서 수행할 행동에 대한 결정은 확률 분포에 따라 확률적으로 샘플링하거나 가장 큰 확률값을 가진 행동을 선택할 수 있다. 전통적 강화학습에서 딥러닝을 기반의 함수 근사화 과정을 적용한 심층강화학습은 최적의 정책을 학습하는 방식에 따라서 가치기반 또는 정책기반 강화학습으로 구분된다. 심층강화학습은 관찰된 환경의 상태 정보 s_t 를 심층신경망에 입력된다. 심층 Q-네트워크(Deep Q-Network : DQN)는 가장 많이 적용되는 모델 중의 하나로서, Q-함수 $Q(s_t,a_t)$ 를 심층 신경망을 통해 근사화하고, 이는 주어진 모든 상태 s_t에 대해 가능한 모든 행동에 대한 큐함수 값을 계산한다. 주어진 정책 π에 대한 가치함수 $Q\pi(s_t,a_t)$는 식 (1)과 같이 정의된다[5].

$$Q\pi(s_t,a_t) = E[R_t \mid s_t,a_t] \qquad (1)$$

ω 는 DQN 의 심층 신경망의 하이퍼 파라미터이며, $Q(s_t,a_t,\omega) \approx Q\pi(s_t,a_t)$ 라고 가정하면, 손실함수를 식 (2) 와 같이 정의할 수 있다. 손실함수를 최소화하면, 이를 토대로 최적의 정책을 구할 수 있다.

$$L(\omega) = E[(r_t + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t))^2] \qquad (2)$$

2.2 심층강화학습 기반 라우팅 기술

심층강화학습은 환경에 대한 모델링 과정이 필요없고, 복잡한 문제를 해결할 수 있기 때문에 라우팅 분야에 적용하는 많은 연구들이 진행되고 있다[9]. 멀티미디어 트래픽에 대해 DRL 을 적용하여 서비스 품질을 최적화하는 라우팅 알고리즘을 제안하였다[10]. 라우팅 최적화를 위해 DDPG 알고리즘을 사용하였고, 에이전트의 목표는 평균 네트워크 지연을 줄이는 것이었다. 14 개 노드로 구성된 네트워크 환경에서 다양한 트래픽 형태에 따라 DRL 에이전트를 학습하고, 성능 분석을 수행하였다. 또한 트래픽 엔지니어링 인식 탐색과 액터-크리틱 기반 우선순위 경험 재생을 통해 지연을 최적화하는 트래픽 제어에 대한 연구가 수행되었다[11]. 심층생성 네트워크를 사용하여 실제 트래픽 데이터에서 QoS 메트릭을 추론하는 연구도 수행되었다[12]. 지식 정의 네트워킹 분야에서, QoS 인식 라우팅 기반의 합성곱 신경망 모델이 제안되었고, 손실과 지연을 QoS 지표로 간주하고 고밀도 신경망을 가진 CNN과 제안된 DDPG의 성능 비교를 제시했다[13]. 또한, [14] 에서는 이중 DQN 모델을 기반으로 우선 순위에 따른 경험 재생을 활용하여 발신지와 목적지 간의 경로를 결정하는 연구를 수행하였다.

III. 시스템 설계 및 문제 정의

3.1 시스템 구성

본 논문에서 고려한 전체 시스템은 통신 서비스를 요청하는 사용자, 라우터들의 상호 결합으로 구성된 데이터 계층 및 SDN 제어 계층으로 구성된다. 그림 2에 SDN 기반 네트워크 구조를 나타내었다. 사용자는 통신 서비스를 이용하는 정보기기를 의미하며, 다양한 QoS 요구 조건을 갖는 응용 서비스를 요청한다. 데이터 계층은 네트워크 장치 간에 데이터 전달 기능을 수행한다. 제어 계층은 라우팅 기능 등을 수행하는 응용 계층과 데이터 계층 간의 통신을 수행하며, 데이터 전달 규칙을 동적으로 업데이트하고 네트워크 자원을 할당, 제어하는 기능을 수행한다. 제어 계층은 현재 네트워크의 모든 상태 정보를 파악하고 있으며, 이를 통해 사용자의 통신 요청이 있을 때 경로 설정 및 각 라우터의 데이터 전달 규칙을 제어한다.

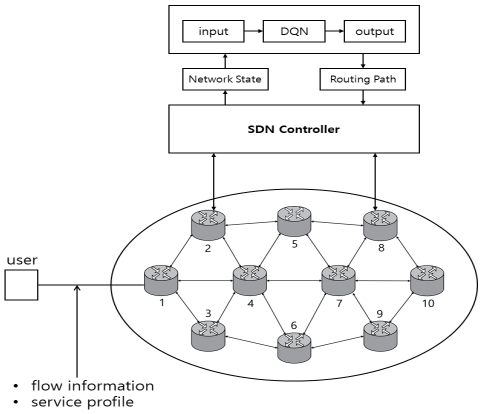

그림 2. SDN 기반의 네트워크 구조

Fig. 2 SDN-based network architecture

네트워크 구성은 양방향 그래프 G(V, E)로 표시되며, V는 모든 라우터의 집합을 나타내고, E는 E = {(i,j)|(i,j) → V × V, i≠j} 와 같은 링크 집합으로 정의된다. 각 링크에 대한 상태 정보는 지연, 손실률, 대역폭에 대해 분리하여 관리한다. QoS 기반 라우팅은 발신지 노드, 목적지 노드, 최소 QoS 요구 파라미터를 입력으로 하여, 해당 조건을 충족하는 경로 리스트를 출력하는 문제로 정의된다[14]. 경로 리스트 내에 포함된 모든 경로는 사용자가 요구하는 최소 대역폭과 최대 지연 및 손실을 만족하여야 한다. QoS 기반 라우팅은 매우 복잡한 문제이므로, 수학적 모델링이 매우 어렵고, 딥러닝을 이용하여 함수 근사화시킬 때에도 학습이 매우 늦어질 수 있다. 따라서, 본 논문에서는 각 흐름이 요구하는 QoS 수준을 유한 개의 서비스 등급으로 구분하고, SDN 네트워크에서 각 링크가 현재 제공할 수 있는 최소 QoS 수준을 미리 계산하여 사용자가 요구하는 QoS 수준을 만족시키는 링크만을 찾아서 발신지에서 목적지까지 경로를 결정하는 문제로 대치될 수 있다[15][16]. 새로 정의된 환경에서 사용자에게 QoS 기반 라우팅 기능을 수행하는 절차는 다음과 같다.

먼저, 통신 서비스를 원하는 사용자는 목적지 정보와 서비스 프로파일을 액세스 라우터에게 전송한다. 서비스 프로파일은 해당 서비스를 제공하기 위해 필요한 대역폭, 지연, 손실률에 대한 최소한의 QoS 파라미터 값을 지정한다. 사용자가 요청한 서비스 프로파일은 대응 함수 또는 변환 테이블을 통해 서비스 등급으로 변환된다. 액세스 라우터는 SDN 제어기를 통해 요구되는 서비스 등급을 충족하는 경로 설정을 요구하는데, 발신지, 목적지 노드 정보 및 서비스 등급 정보를 포함하여 전송한다. SDN 라우팅 계층은 동일한 대응 함수를 이용하여 현재 네트워크 상태 정보를 각 링크들이 제공하는 서비스 등급의 행렬로서 변환하며, 이를 DQN 의 상태 정보를 간주한다. DQN 에이전트는 발신자, 목적지 및 서비스 등급을 입력받고, 환경과 상호 작용을 통해 서비스 등급을 만족시키는 경로 정보를 결정한다. DQN 에이전트는 경로 리스트에 대한 정보를 SDN 제어기를 통해 각 라우터의 데이터 전달 규칙에 적용하고, 사용자에게 데이터 전송 시작을 승인한다.

Ⅳ. DRL 기반 라우팅 알고리즘

본 장에서는 서비스 프로파일에 지정된 QoS 파라미터를 만족하면서 통신을 수행하는 두 노드 간의 경로를 찾기 위한 DQN 기반 라우팅 알고리즘에 대해 설명한다.

4.1 상태 공간

상태 공간은 네트워크의 현재 상태 정보를 포함해야 하는데, SDN 제어기를 통해 일차적으로 네트워크의 연결 정보 및 링크 상태 정보를 파악한다. 초기 상태 정보는 대역폭, 지연, 손실률 등 QoS 파라미터에 대해 |V| × |V| 크기의 2차원 행렬을 이용하여 표현한다. 각 파라미터의 값은 단위가 다르고, 스케일이 다를 수 있기 때문에 정규화 과정을 거친다. 이후 대응 함수를 통해 각 링크 별 QoS 파라미터 정보로부터 해당 링크의 서비스 등급값으로 변환된 통합된 2차원 행렬 형태로 변환되고, 변환된 행렬 정보를 현재 네트워크 상태 정보로 간주한다.

4.2 행동 공간

강화학습에서 행동 공간은 각 상태마다 에이전트가 선택할 수 있는 행동들의 집합을 의미한다. 라우팅 문제에서의 행동 공간은 경로상의 노드에 연결된 링크 중의 하나를 선택해야 하기 때문에 각 중간 노드 마다 선택할 수 있는 행동 공간이 달라지게 된다. 하지

만, 강화학습 모델에서 모든 상태의 행동 공간은 동일해야 한다. 각 노드에서 선택할 수 있는 행동은 해당 노드에서 나가는 출력 링크로 제한되지만, 강화학습 모델에 부합하고, 에이전트 구현을 단순화하기 위해 모든 링크를 선택 가능하도록 한다. 에이전트가 선택한 링크에 대한 적절성에 대한 판단은 환경이 수행하며, 양 또는 음의 보상값을 전달함으로써, 에이전트가 해당 노드에 연결된 출력 링크를 적절한 행동으로 결정하도록 학습을 시킨다. 따라서, 행동 공간은 네트워크에 존재하는 모든 링크의 집합으로 정의된다. 즉, 행동 공간 벡터는 A = [a_1, a_2,. . . , $a_{|E|}$] 이며, 각각의 행동은 네트워크의 링크 (i, j) ∈ E 에 해당한다.

4.3 보상 함수

보상 함수는 상태변환 확률과 함께 환경이 관리하는 모델이며, 에이전트가 보상값을 통해 최적의 정책을 학습하게 된다. 따라서, DRL 알고리즘의 학습 정확도를 높이기 위해서는 보상 함수를 적절하게 지정하는 것이 매우 중요하다. 특히, 에이전트의 행동 공간이 네트워크에 연결된 어떠한 링크라도 선택이 가능하도록 정의하였기 때문에 다양한 상황을 고려하여 보상함수를 설정할 필요가 있다. 일차적으로 식(3)과 같이 에이전트가 취한 행동에 대한 유효성 여부를 판단하여 보상값을 반환한다. 선택한 행동이 유효하지 않으면 에이전트는 −|V|/2 의 음의 보상값을 받는다. 유효한 행동을 선택했을 경우에 별도로 정의된 함수 $g((i, j))$ 에 의해서 보상값을 결정한다[14]. 에이전트는 일정한 타임 스텝 내에서 발신지 노드에서 목적지 노드까지의 경로 리스트를 찾아야 한다. 최대 타임 스텝을 초과하였을 경우 음의 보상값을 반환하여야 한다. 임의의 노드 z에서 에이전트는 링크 (i, j)에 해당하는 시간 단계 t에서 행동을 선택하면, 환경은 보상함수 $f((i, j))$에 의해 보상값 r_t를 결정하여, 다음 상태 정보와 함께 에이전트에 전달한다

$$r_t = f((i,j)) \qquad (3)$$

$$f((i,j)) = \begin{cases} g((i,j)) & (i,j) \in \varepsilon_{valid}^{z} \\ -\dfrac{|v|}{2} & otherwise \end{cases}$$

여기서 ϵ_{valid}^{z}는 노드 z의 유효한 행동 집합을 나타낸다. 즉, 노드 z에서 출력 링크를 선택할 경우에만 유효한 행동으로 간주한다. 유효한 행동을 선택하였을 때 해당 행동의 대응되는 링크가 사용자가 요구한 QoS 등급을 만족하면 양의 보상값을, 충족하지 못하면 0 의 보상값을 반환한다. 이를 위해 이미 선택된 노드를 향하는 링크를 선택할 경우 매우 큰 음의 보상을 지정한다. 이를 위해 $\epsilon_{visited}$ 리스트는 각 에피소드의 시작 부분에 비어있는 것으로 정의되고, 이후 링크로 채워지는 집합을 나타낸다. 에이전트가 에피소드에서 동일한 링크를 반복적으로 선택하여 네트워크 루프에 갇히지 않는다. 선택한 링크에 대상 노드가 목적지 노드 y 일 경우에는 이 상태가 터미널 상태이며, 환경은 에이전트에게 가장 높은 양의 보상값 |V| 를 반환한다. 에이전트가 액션 공간을 탐색하는 동안 무한 루프에 갇히지 않도록 각 에피소드를 T 타임 스텝으로 제한한다. 현재의 타입 스텝이 총 에지 수보다 크면 에이전트가 음의 보상값을 받게 되고, 에피소드는 -|V| 의 높은 패널티로 종료된다. 나머지 경우는 에이전트가 선택한 행동에 대한 링크가 요구되는 QoS 수준과 비교하여 적절한 양의 보상과 음의 보상을 반환한다.

4.4 DQN 기반 라우팅 알고리즘 구현

정의한 상태 공간, 행동 공간, 보상 함수를 DQN에 적용하였다. 입력 계층은 네트워크의 모든 링크에 대한 서비스 등급값으로 변환된 2차원 행렬 정보를 |V| × |V| 크기의 1차원으로 차원이 변환된 정보이다. DQN 의 신경망 모델은 K 개의 뉴런으로 구성된 3 개의 은닉 계층으로 구성되며, 활성화 함수로 ReLU를 적용하였다. 출력 계층의 크기는 행동 공간의 크기 |A| 로 지정하였고, 활성화 함수는 각 행동에 대한 큐 함수값을 출력하도록 선형함수를 적용하였다. 에이전트는 출력된 큐함수 값 중에서 가장 큰 큐함수 값을 나타내는 행동을 선택한다. 또한 DRL 에이전트의 훈련 과정에서 배치 크기는 64, 학습률은 0.001, 버퍼 크기 는 5000 으로 지정하였다. 에이전트가 학습을 위해 입실론 그리디 방식의 정책을 선택하였고, 랜덤 정책이 선택될 확률을 나타내는 초기 입실론 값은 0.9 로, 학습이 진행됨에 따라 랜덤 정책을 선택할 확률을 낮추기 위한 입실론 감쇄값은 0.99 로 각각 설정하였다. 라우팅을 학습하기 위한 DQN 알고리즘

은 그림 3에 나타내었다. 알고리즘 초기화 과정에서 네트워크 노드 수와 링크 정보를 지정하여 환경의 인스턴스를 만든다. 이후, 재생 버퍼, 메인 Q-네트워크 및 타겟 Q-네트워크의 하이퍼 파라미터를 초기화된다.

알고리즘은 총 N 개의 에피소드에 대해 실행된다. 각 에피소드가 시작될 때 네트워크 링크 상태 정보는 정규화된다. 대응 함수를 통해 각 링크 별 QoS 파라미터 정보로부터 해당 링크의 서비스 등급값으로 변환된 통합된 2차원 행렬 형태로 변환된다. 이미 방문한 에지에 대한 보상함수를 알리는 데 사용되는 빈 세트 $\epsilon_{visited}$가 생성된다. 각 에피소드에는 T 시간 단계의 기간이 있다 (여기서 T = | E |).

Initialize Environment
Initialize replay memory
Initialize main, target deep-Q network with weights θ
for episode = 1 to N **do**
  Reset edge' s metrics value
  Normalize each metric value
  Get source-destination pair (x, y) info. from agent
  Create state space s_t
  Create an empty set $\varepsilon_{visited}$
  for t = 1 to T **do**
    With probability ϵ select random action a_t
    Otherwise select $a_t = \{\arg\max Q(s,a;\theta)\}$
    Get valid actions set ε^x_{valid}
    if $a_t \in \varepsilon^x_{visited}$ **then**
      Execute action a_t in the environment
      Obtain reward r_t a
      Update $\varepsilon_{visited}$
      $x = z$ (where z is the new selected node)
      Update state space s_{t+1}
    else
      Obtain reward r_t
      $s_{t+1} = s_t$
    end
    Store transition (s_t, a_t, r_t, s_{t+1}) in replay memory
    Sample random mini-batch of transitions
      (s_j, a_j, r_j, s_{j+1}) from replay memory
    Set y_j^{DQN} using eqn.
    Perform gradient descent step
      $(y_j^{DQN} - Q(s_j, a_j; \theta_t, \eta_t, \zeta_t))^2$ w.r.t θ
    Update target network weights every τ time steps
    if $x == y$ **then**
      break
  end
end

그림 3. 경로 결정을 위한 DQN 학습 알고리즘
Fig. 3 DQN training algorithm for path determination

각 시간 단계 t에서 엡실론 그리디 방식을 사용하여 행동 at가 선택된다. 선택된 동작이 유효한 동작이면 환경에서 실행되고, 보상함수에 따라 보상 rt를 얻고 $\epsilon_{visited}$에 at를 포함하고 상태 공간 벡터를 업데이트한다. 그렇지 않으면 선택한 행동이 유효한 행동이 아닌 경우 음의 보상값을 전달한후 다음 반복을 위해 동일한 상태 공간을 변경하지 않고 사용한다. st, at, rt, st + 1을 얻은 후 전환은 경험 재생 버퍼에 저장된다. 재생 버퍼에 일정 이상의 전환 정보가 축적되면, 재생 버퍼에서 임의의 미니 배치 전환을 샘플링하고 Θ에 대한 경사 하강법을 사용하여 심층 신경망의 가중치를 최적화하여 손실을 최소화한다. 일정한 주기(τ)마다 메인 네트워크 파라미터를 통해 타켓 네트워크 파라미터를 갱신한다. 에피소드의 반복은 목적지 노드까지 경로가 완성되거나, t > |E| 인 경우 종료한다. 제안한 알고리즘의 검증을 위한 네트워크 구조는 그림 4와 같다. 10 개 노드로 구성된 네트워크 구조에 대해서 각 링크별 QoS 파라미터를 지정하였다. 각 링크가 가질 수 있는 지연, 대역폭 및 손실률 값은 각각 (1, 100) ms, (50, 100) Mbps, (0.01, 1) 에서 균등하게 선택되었다. 네 개의 QoS 단계에 대해 노드 1 에서 노드 10 으로 향한 경로 설정을 적절하게 학습하는 가에 대해 분석하였다.

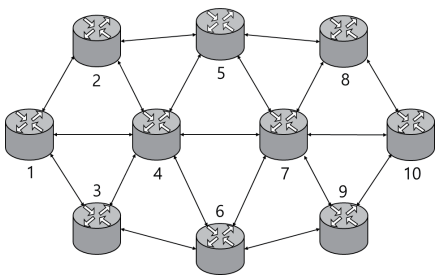

그림 4. 알고리즘 검증을 위한 네트워크 구조
Fig. 4 Network topology for verification of algorithm

알고리즘의 검증을 위해 고정된 네트워크 환경에서 학습을 수행하고, 매 타임 스텝마다 발신지와 목적지 간에 선택한 경로 리스트를 출력하고, 하나의 에피소드가 완료되었을 때 누적된 보상값을 출력하도록 하였다. 학습 초기에는 상당 기간 환경을 탐색하는 과정을 수행하기 때문에 에이전트는 유효하지 않거나 이

미 선택된 행동을 다시 선택하게 된다. 대부분 에피소드가 비정상으로 종료되고, 따라서 매우 큰 누적된 음의 보상값을 받은 것으로 나타났다. 에이전트가 점차 학습이 진행하게 되면, 유효하지 않은 행동과 네트워크 루프를 피하는 행동을 선택함에 따라 받게 되는 누적된 음의 보상값이 점점 감소하는 추세를 나타내었고, 각 타임 스텝마다 올바른 경로를 선택함을 확인하였다. 제한된 네트워크 환경에 대해 본 알고리즘을 검증하였다. 다양한 네트워크 상태에 대해서 임의의 두 쌍의 노드 간의 라우팅 기능에 대한 성능을 평가 중에 있다. 추가적으로 기존 라우팅 방식과의 작업의 복잡도와 성능 정확도에 대한 성능 비교 연구를 수행할 예정이다.

Ⅳ. 결 론

본 논문은 소프트웨어 정의 네트워크(SDN)에서 심층강화학습을 활용하여 서비스 품질을 고려하여 최적의 경로를 결정하는 라우팅 알고리즘을 제안하였다. 학습을 위한 심층강화학습 모델은 DQN 을 기반으로 하고, 지연, 대역폭, 손실률 등 QoS 파라미터를 기준으로 최적의 경로를 선택하는 라우팅 기능을 학습한다. 본 논문에서는 각 흐름이 요구하는 QoS 수준을 단계화하고, SDN 네트워크에서 각 링크가 현재 제공할 수 있는 최소 QoS 수준을 미리 계산하여 흐름이 요구하는 QoS 수준을 만족시키는 링크만을 찾아서 발신지에서 목적지까지 경로를 결정함으로써 계산 및 학습의 복잡도를 완화시킬 수 있다. 시뮬레이션 결과는 제안한 알고리즘이 일정한 에피소드를 진행하게 되면 올바른 경로를 선택하게 되고, 학습이 성공적으로 수행됨을 확인하였다. 향후 다양한 네트워크 상태에 대해 검증이 필요하며, 기존 라우팅 방식과의 작업의 복잡도와 성능 정확도에 대한 성능 비교가 필요하다.

감사의 글

“이 논문은 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구 사업임. (No. 2019R1I1A3A0106291)“

References

[1] N. Feamste, J. Rexford, and E. Zegura, “The Road to SDN: An Intellectual History of Programmable Networks,” *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 2, 2014, pp. 87 – 98.

[2] A. Yassine, H. Rahimi, and S. Shirmohammadi, ‘‘Software defined network traffic measurement: Current trends and challenges,’’ *IEEE Instrum. Meas. Mag.*, vol. 18, no. 2, 2015, pp. 42 – 50.

[3]. W. Guck, A. V. Bemten, M. Reisslein, and W. Kellerer, “Unicast QoS routing algorithms for SDN: A comprehensive survey and performance evaluation,” *IEEE Communications Surveys Tutorials,* vol. 20, no. 1, 2018, pp. 388 – 415.

[4] M. Karakus and A. Durresi, “Quality of service in software defined networking: A survey,” *Journal of Network and Computer Applications,* vol. 80, 2017, pp. 200 – 218.

[5] V. Mnih, K. Kavukcuoglu and D. Silver, “Human-level control through deep reinforcement learning,” *Nature,* vol. 518, no. 7540, 2015, pp. 529 - 535.

[6] N. C. Luong et al., "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," i*n IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, 2019, pp. 3133-3174

[7] S.-C. Lin, I. F. Akyildiz, P. Wang, and M. Luo, “QoS-aware adaptive routing in multi-layer hierarchical software defined networks: A rein- forcement learning approach,” i*n Services Computing (SCC), 2016 IEEE International Conference on. IEEE*, 2016, pp. 25 – 33.

[8] G. Stampa, M. Arias and A. Cabellos, “A deep-reinforcement learning approach for software defined networking routing optimization,” arXiv preprint arXiv:1709.07080, 2017.

[9] X. Huang, T. Yuan, G. Qiao and Y. Ren, “Deep reinforcement learning for multimedia traffic control in software defined networking,“ *IEEE Network*, vol. 32, no. 6, 2018, pp. 35 - 41.

[10] B. Guo, X. Zhang, Y. Wang, and H. Yang, “Deep Q-network based multimedia multi -service QoS optimization for mobile edge computing systems,“ *IEEE Access,* vol. 7, 2019, pp. 160961 - 160972.

[11] A. Valadarsky, M. Schapira, and A. Tamar, “Learning to route with deep reinforcement learning,” *in NIPS Deep Reinforcement Learning Symposium*, 2017.

[12] Xu, J. Tang, C. Yin, Y. Wang, and G. Xue, “Experience-driven congestion control: When multi-path tcp meets deep reinforcement learning,” *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, 2019, pp. 1325 - 1336.

[13] S. Xiao, D. He, and Z. Gong, “Deep-q: Traffi c-driven qos inference using deep generative network,” *in Proceedings of the 2018 Workshop on Network Meets AI & ML*, 2018, pp. 67 - 73.

[14] S. Q. Jalil, M. Husain Rehmani and S. Chalup, "DQR: Deep Q-Routing in Software Defined Networks," *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1-8.

[14] S. Q. Jalil, M. Husain Rehmani and S. Chalup, "DQR: Deep Q-Routing in Software Defined Networks," *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1-8.

[15] S. Lee, “Design and Application of LoRa-based Network Protocol in IoT Networks,” *J. of the Korea Institute of Electronic Communication Sciences,* vol. 14, no. 6, 2019, pp. 1089-1096.

[16] S. Jung and Lee, “A Queue Management Mechanism for Service groups based on Deep Reinforcement Learning,” *J. of the Korea Institute of Electronic Communication Sciences,* vol. 15, no. 6, 2020, pp. 1099-1104.

저자 소개

이성근(Sung-Keun Lee)

1985년 고려대학교 전자공학과 졸업(공학사)
1987년 고려대학교 대학원 전자공학과 졸업(공학석사)
1995년 고려대학교 대학원 전자공학과 졸업(공학박사)
1987년 ~ 1992년 : 삼성전자 정보통신연구소
1996년 ~ 1997년 : 삼성전자 네트워크 연구팀
2017년 ~ 2018년 : Georgia Institute of Technology ECE 방문교수
1997년 ~ 현재 순천대학교 멀티미디어공학과 교수
※ 관심분야 : 심층강화학습 기반 QoS 보장 기술, AI 기반 태양광 예측 시스템, 스마트 농업, 멀티미디어 통신