(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0039772
(43) 공개일자 2025년03월21일

(51) 국제특허분류(Int. Cl.)
*G06N 5/04* (2023.01) *G06N 3/092* (2023.01)
*G06N 7/01* (2023.01)
(52) CPC특허분류
*G06N 5/04* (2023.01)
*G06N 3/092* (2023.01)
(21) 출원번호 10-2023-0122651
(22) 출원일자 2023년09월14일
심사청구일자 2023년09월14일

(71) 출원인
**고려대학교 산학협력단**
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
**백상헌**
서울특별시 서초구 고무래로 35, 109동 1004호
**고한얼**
세종특별자치시 달빛로 80, 1216동 1903호
**정현재**
경기도 수원시 권선구 당진로14번길 37, 101동 1310호
(74) 대리인
**김등용**

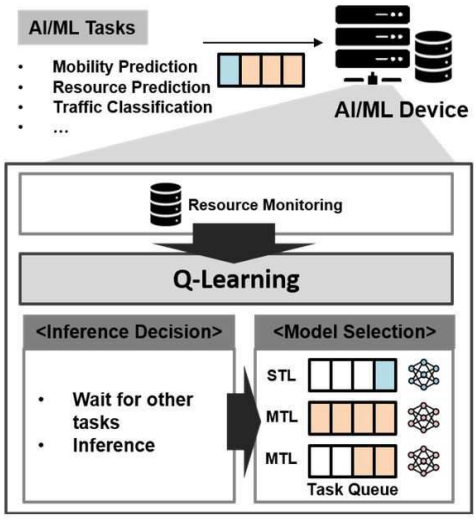
전체 청구항 수 : 총 9 항

(54) 발명의 명칭 **태스크 요청 속도를 고려한 추론 모델 선택 방법**

## (57) 요 약

태스크 요청 속도를 고려한 추론 모델 선택 방법 및 장치가 개시된다. 상기 추론 모델 선택 방법은 적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되고, 상기 컴퓨팅 장치의 컴퓨팅 자원을 모니터링하는 단계, 태스크를 수신하는 단계, 수신된 태스크에 대한 추론을 수행할 추론 모델을 선택하는 단계, 태스크를 선택된 추론 모델의 큐(queue)에 입력하는 단계, 및 추론 동작을 수행하는 단계를 포함하한다.

*대 표 도* - 도1



(52) CPC특허분류
*G06N 7/01* (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193180
과제번호 2022-0-01015-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 6G핵심기술개발
연구과제명 지능형 6G 모바일 코어 네트워크 후보 요소기술 개발
기 여 율 1/1
과제수행기관명 고려대학교 산학협력단
연구기간 2023.01.01 ~ 2023.12.31

# 명 세 서

## 청구범위

### 청구항 1

적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되는 추론 모델 선택 방법에 있어서,

상기 컴퓨팅 장치의 컴퓨팅 자원을 모니터링하는 단계;

태스크를 수신하는 단계;

수신된 태스크에 대한 추론을 수행할 추론 모델을 선택하는 단계;

태스크를 선택된 추론 모델의 큐(queue)에 입력하는 단계; 및

추론 동작을 수행하는 단계를 포함하는 추론 모델 선택 방법.

### 청구항 2

제1항에 있어서,

상기 모니터링하는 단계는 미리 정해진 주기에 따라 주기적으로 수행되는,

추론 모델 선택 방법.

### 청구항 3

제2항에 있어서,

상기 추론 모델을 선택하는 단계는, 상기 미리 정해진 주기 동안에 수신된 태스크에 대하여 추론 모델을 선택하고,

추론 모델은 적어도 하나의 STL(Single-task Learning) 모델과 적어도 하나의 MTL(Multi-task Learning) 모델을 포함하는,

추론 모델 선택 방법.

### 청구항 4

제3항에 있어서,

상기 추론 모델을 선택하는 단계는,

MDP(Markov Decision Process) 모델을 큐러닝(Q-Learning)으로 변형한 결과로 도출된 큐함수를 풀이하고,

상기 큐함수는 $Q_*(S,A) = \max_\pi[Q_\pi(S,A)]$이고,

상기 S는 상태 공간(State Space)이고, 상기 A는 액션 공간(Action Space)이고, 상기 π는 정책(policy)인,

추론 모델 선택 방법.

### 청구항 5

제4항에 있어서,

상기 추론 동작을 수행하는 단계는,

선택된 모델이 STL 모델인 경우, 큐에 입력된 태스크가, 대기 시간 없이 STL 모델에 입력되어 추론 동작이 수행되는,

추론 모델 선택 방법.

### 청구항 6

제5항에 있어서,

상기 추론 동작을 수행하는 단계는,

선택된 모델이 MTL 모델인 경우, 큐가 태스크로 가득 찬 경우에 큐에 입력된 태스크가 MTL 모델로 입력되어 추론 동작이 수행되는,

추론 모델 선택 방법.

**청구항 7**

제5항에 있어서,

상기 추론 동작을 수행하는 단계는,

선택된 모델이 MTL 모델인 경우, 큐에 저장되어 있는 적어도 하나의 태스크가 수신된 시점으로로부터 경과된 시간과 추론에 필요한 컴퓨팅 시간의 합이 요구 지연 시간으로부터 미리 정해진 시간을 뺀 값과 같을 때, 큐에 저장되어 있는 모든 태스크가 MTL 모델로 입력되어 추론 동작이 수행되는,

추론 모델 선택 방법.

**청구항 8**

제5항에 있어서,

상기 추론 동작을 수행하는 단계는,

선택된 모델이 MTL 모델인 경우, 태스크가 큐에 대기한 시간과 추론에 필요한 컴퓨팅 시간의 합이 요구 지연 시간으로부터 미리 정해진 시간을 뺀 값과 같을 때, 큐에 저장되어 있는 모든 태스크가 MTL 모델로 이력되어 추론 동작이 수행되는,

추론 모델 선택 방법.

**청구항 9**

제8항에 있어서,

추론 결과를 송신하는 단계를 더 포함하는,

추론 모델 선택 방법.

## *발명의 설명*

### *기 술 분 야*

[0001] 본 발명은 태스크 요청 속도에 따라 단일 태스크 러닝(Single-task Learning, STL) 모델 또는 다중 태스크 러닝(Multi-task Learning, MTL) 모델을 선택하는 방법에 관한 것이다.

### *배 경 기 술*

[0002] 딥러닝(Deep Learning, DL)이란 데이터를 기반으로 학습하고, 이후 추론하는 과정을 통해 분석 결과를 도출하는 기술이다. 일반적으로 딥러닝은 단일 태스크를 수행하는 뉴럴 네트워크 즉, 단일 태스크 러닝(Single-task Learning, STL) 모델을 사용하는 것을 의미하며, 학습 단계에서 네트워크를 구성하는 뉴럴의 가중치를 계산하는 방법을 통해 데이터를 학습한다. 따라서 기존의 STL은 단일 태스크에 초점을 맞추었기 때문에, 다중 태스크를 처리해야 하는 환경에서는 컴퓨팅 자원 측면에서의 제약이 있었다. 예를 들어, 모바일 트래픽 특성 분석을 위해 짧은 시간 이내에 모바일 트래픽 분류, 모바일 트래픽 예측, 모바일 이동성 예측, RAN의 자원 사용량 예측 등에 대한 분석 결과가 필요한 경우, STL 모델만 사용하면 태스크 수행을 위해 각 태스크에 컴퓨팅 자원을 할당해야 하고 이는 많은 양의 컴퓨팅 자원을 소비한다. 이와 같이 한정된 시간 내에 여러 개의 태스크 요청이 있는 경우 추론 가능한 네트워크 장치(e.g., 모바일, 엣지 장비, 코어 장비)에서 높은 컴퓨팅 부하로 인해 태스크의 요구 지연 시간을 만족하지 못할 수 있다.

[0003] 다중 태스크 러닝(Multi-task Learning, MTL) 모델은 다중 태스크를 하나의 모델로 처리하기 위해 개발되었다.

다중 태스크 모델은 여러 태스크를 동시에 학습하고 추론할 수 있으며, 각 태스크에 대한 독립적인 특징을 추출한다. 초기에 MTL 모델은 태스크 간 데이터 공유 및 비교를 통해 태스크 간의 상호 의존성을 최소화하고 모델 정확도를 높이는 목적으로 고안되었으나, 최근 MTL 모델이 학습 시간과 컴퓨팅 자원 사용을 줄일 수 있다는 장점을 활용한 연구가 여러 분야에서 활발히 진행되고 있다. 이와 같이 MTL 모델은 STL 모델보다 컴퓨팅 자원 측면에서 효율적인 결과를 얻을 수 있다는 장점이 있으나, 각 태스크가 동시에 컴퓨팅 되므로 태스크의 각기 다른 요구 지연 시간을 만족하지 못할 수 있다는 문제가 있다. 따라서 MTL 모델의 장점을 활용하기 위해 태스크 요청 속도와 태스크의 요구 지연 시간과 같은 각 태스크의 특성을 고려하고, 이에 따라 STL 모델과 MTL 모델을 선택적으로 사용하는 것이 중요하다.

[0004] 기존의 딥러닝 관련 기법들은 단일 태스크를 기준으로 한다. 이는 태스크 요청 속도와 태스크를 처리하는 데에 소비되는 컴퓨팅 자원을 고려하지 않고 있기 때문에 태스크 요청 속도가 높은 환경에서는 자원 부족으로 인해 원하는 추론 지연 시간을 얻기 힘들다. 따라서 본 발명에서는 태스크 요청 속도와 컴퓨팅 자원을 고려하면서 태스크 요구 지연 시간을 만족하기 위해 적합한 추론 모델을 선택하는 방안을 제안하고자 한다.

[0005] MTL 모델은 정확성을 유지하기 위해 일반적으로 STL 모델과 같거나 더 높은 컴퓨팅 자원(FLOPS)을 소비한다. 따라서 본 발명에서는 MTL 모델이 STL 모델과 같은 컴퓨팅 자원을 소비하는 것을 가정한다. 태스크 종류는 모바일 트래픽 분류, 모바일 트래픽 예측, 모바일 이동성 예측, RAN의 자원 사용량 예측 등을 예시로 들 수 있으며, 각 네트워크 태스크는 초에서 분 단위의 넓은 범위 안에서의 요구 지연 시간을 가진다. 추론 가능한 네트워크 장치(e.g., 모바일, 엣지 장비, 코어 장비)는 태스크 요청을 받으면 STL 모델을 사용하여 바로 추론할지 혹은 MTL 모델로 함께 추론할 태스크를 기다릴지 결정한다.

[0006] 태스크 요청 속도가 낮은 경우 모든 태스크에 대해 STL을 선택해도 태스크 요구 지연 시간을 만족시킬 수 있으나, 태스크 요청 속도가 높은 경우 모든 태스크에 대해 STL을 선택한다면 컴퓨팅 자원 제약으로 인해 컴퓨팅 자원 효율이 떨어질 뿐만 아니라, 태스크의 요구 지연 시간을 만족하지 못할 수 있다. 마찬가지로 태스크 요청 속도가 높은 경우 모든 태스크에 대해 MTL을 선택한다면 컴퓨팅 자원 효율성은 높으나 각 태스크의 요구 지연 시간을 만족하지 못할 수 있다. 이를 위해 본 발명에서는 태스크 요청 속도 변화에 따라 추론 모델을 적응적으로 선택하는 방안을 제시한다. 즉, 태스크 요청 속도에 따라 STL 모델과 MTL 모델을 적응적으로 선택하여 컴퓨팅 자원을 효율적으로 활용하면서 태스크의 요구 지연 시간을 만족할 수 있다.

[0007] 이를 위해 본 발명에서는 태스크 특성(e.g., 태스크 요청 속도 및 각 태스크의 요구 지연 시간)을 고려하여 효율적으로 컴퓨팅 자원을 활용하기 위한 추론 모델을 선택하는 방법을 제안한다. 제안 방법에서는 태스크 특성에 따라 최적의 추론 모델을 결정하기 위해 마르코브 결정 프로세스(Markov Decision Process, MDP)로 문제를 공식화하고, 큐러닝(Q-Learning)으로 최적의 정책(policy)을 찾는다.

## 선행기술문헌

### 특허문헌

[0008] (특허문헌 0001) 대한민국 공개특허 제2023-0029468호 (2023.03.03. 공개)

(특허문헌 0002) 대한민국 공개특허 제2021-0091919호 (2021.07.23. 공개)

(특허문헌 0003) 대한민국 공개특허 제2023-0032690호 (2023.03.07. 공개)

(특허문헌 0004) 대한민국 공개특허 제2023-0068989호 (203.05.18. 공개)

## 발명의 내용

### 해결하려는 과제

[0009] 본 발명이 이루고자 하는 기술적인 과제는 태스크 요청 속도를 고려한 추론 모델 선택 방법을 제공하는 것이다.

### 과제의 해결 수단

[0010] 본 발명의 일 실시예에 따른, 태스크 요청 속도를 고려한 추론 모델 선택 방법 및 장치가 개시된다. 상기 추론 모델 선택 방법은 적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되는 추론 모델 선택 방법으

로써, 상기 컴퓨팅 장치의 컴퓨팅 자원을 모니터링하는 단계, 태스크를 수신하는 단계, 수신된 태스크에 대한 추론을 수행할 추론 모델을 선택하는 단계, 태스크를 선택된 추론 모델의 큐(queue)에 입력하는 단계, 및 추론 동작을 수행하는 단계를 포함한다.

## 발명의 효과

[0011] 본 발명의 실시예에 따른, 태스크 요청 속도를 고려한 추론 모델 선택 방법에 의할 경우, 태스크 요청 속도와 태스크의 요구 지연 시간을 고려하여 각 태스크에 적합한 추론 모델을 선택할 수 있는 효과가 있다.

## 도면의 간단한 설명

[0012] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 본 발명의 일 실시예에 따른, 추론 모델 선택 장치를 도시한다.

도 2는 태스크 요청 속도가 1일 때, 제안 기법과 비교 기법의 평균 보상 변화를 나타낸다.

도 3은 태스크 요청 속도가 2일 때, 제안 기법과 비교 기법의 평균 보상 변화를 나타낸다.

도 4는 도 1에 도시된 추론 모델 선택 장치에 의해 수행되는 추론 모델 선택 방법을 설명하기 위한 흐름도이다.

## 발명을 실시하기 위한 구체적인 내용

[0013] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.

[0014] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0015] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.

[0016] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0017] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0018] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0019] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시예들을 상세히 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부

재를 나타낸다.

[0020] 도 1은 본 발명의 일 실시예에 따른 추론 모델 선택 장치를 도시한다. 추론 모델 선택 장치는 적어도 프로세서(Processor) 및/또는 메모리(memory)를 포함하는 컴퓨팅 장치로 구현될 수 있다. 따라서, 추론 모델 선택 방법을 이루는 단계들 중 적어도 일부는 컴퓨팅 장치에 포함되는 프로세서의 동작으로 이해될 수도 있다. 컴퓨팅 장치는 PC(Personal Computer), 서버(server), 랩탑 컴퓨터, 태블릿 PC, 스마트폰(smart phone) 등을 포함할 수 있다.

[0021] 실시예에 따라, 추론 모델 선택 장치는 추론 수행 가능한 추론 장비(AI/ML Device)와 해당 장비에 요청되는 AI/ML 태스크로 이루어진다. 추론 장비는 매 시간 τ 마다 컴퓨팅 자원 상태를 확인하여, 태스크 요청 시 추론 모델을 선택하는 데에 남은 컴퓨팅 자원을 고려할 수 있도록 한다. 추론 장비는 태스크 추론을 위해 여러 STL, MTL 모델을 최신으로 유지하고, 각 모델에 대해 태스크 큐를 유지한다. 태스크가 요구 지연 시간에 대한 정보와 함께 추론 장비에 요청되면, 추론 장비는 유지하고 있는 STL, MTL 모델로 추론 가능한 태스크인지 먼저 확인하고, 이를 큐러닝 모듈에 입력한다. 큐러닝 모듈에서 내린 추론 모델 결정에 의해 태스크는 STL 모델 큐 혹은 MTL 모델 큐에 이력된다. STL 모델 큐에 입력된 태스크의 경우 대기 시간 없이 바로 추론이 시작되며, MTL 큐에 입력된 태스크의 경우 큐가 다 차거나 큐에 대기하는 태스크 중 요구 지연 시간에 근접한 태스크가 있다면 추론된다.

[0022] 상술한 정보(추론 장비의 가용 자원 정보, 태스크의 요구 지연 시간 정보)들은 큐러닝 모듈의 입력(input)으로 입력된다. 해당 기능에서는 후술하는 동작 절차를 수행한 후 출력(output)으로 태스크의 추론 모델 선택을 하게 된다. 추론 장비에서 추론 모델 선택 결정은 시스템 모델 수식화를 위해, 태스크는 전체 N 개까지 존재하고 각각 다른 요구 지연 시간과 태스크 경과 시간(큐에 대기 시간 및 컴퓨팅 시간(추론 시간을 의미할 수 있음))을 가지는 것으로 정의한다. 추론 장비에서의 추론 모델 선택은 시간 τ 마다 장비에 존재하는 모든 태스크에 대해 동시에 이루어진다. 선택은 각 태스크의 요구 지연 시간을 만족하면서 자원 사용을 최소화하는 것을 목적으로 한다.

[0023] 동작 절차

[0024] 동작 과정 1 : MDP 모델링

[0025] 상태 공간(State Space) S 정의

[0026] 상태 공간 S는 아래 수학식과 같이 정의된다.

[0027] $$S = C \times \prod_{n}^{N} K_n \times T_n$$

[0028] 상기 수학식에서, $C=\{0,1,2,\cdots,C_{max}\}$는 컴퓨팅 자원을, $n=\{0,1,2,\cdots,N\}$은 태스크 개수를 나타낸다. 각 태스크는 각기 다른 요구 지연 시간을 가질 수 있다. $K=\{0,1,2,3\}$는 태스크 상태를 나타낸다. 구체적으로, $K_n=0$은 태스크 요청이 없는 상태를, $K_n=1$은 태스크 요청 도착 및 대기인 상태를, $K_n=2$는 태스크 추론 중인 상태를, $K_n=3$은 태스크 추론 완료된 상태를 나타낸다. $T=\{0,1,2,\cdots,T_{max}\}$는 태스크의 경과 시간을 나타낸다.

[0029] 액션 공간(Action Space) A 정의

[0030] 액션 공간은 다음 수학식과 같이 정의된다.

[0031] $$A = \prod_{n}^{N} A_n$$

[0032] 상기 수학식에서, $A_n=\{0,1,2,3,4\}$이고, $A_n=0$은 태스크 n에 대해 추론을 수행하지 않고 대기함을 의미한다. $A_n=1$은 STL 모델로 추론함을 의미한다. $A_n=2$, $A_n=3$, $A_n=4$은 각각 2개, 3개, 4개 태스크를 출력으로 하는 MTL 모델로 추론함을 의미한다.

[0033] 보상 함수(Reward Function) R 정의

[0034] 보상 함수는 다음 수학식과 같이 정의된다.

[0035] $$R = \frac{t}{r}$$

[0036] 상기 수학식에서, r={0,1,…,$r_{max}$}이고, 추론 횟수를 나타낸다. t={0,1,2,…,$t_{max}$}이고, 출력 태스크의 개수를 나타낸다. 태스크의 요구 지연 시간을 만족하면 R, 그렇지 않으면 0을 얻도록 한다.

[0037] 동작 과정 2 : 최적화 과정

[0038] 동작 과정 1에서 정의한 MDP 모델의 최종 목적은 특정 상태(state)에서 특정 액션(action)을 취하는 최적의 정책(policy)을 찾는 것이다. 위 환경의 경우, 모든 상태, 즉 모든 경우의 컴퓨팅 자원과 태스크 상태 및 태스크의 지연 시간에 대해 어떤 액션을 취했을 때 결과적으로 가장 큰 보상을 얻을 수 있는지에 대한 경우의 수가 매우 많기 때문에 이를 정확히 알기 어렵다. 따라서, 본 발명에서는 환경에 대한 모든 설명 없이도 최적의 정책을 탐색할 수 있는 큐러닝을 이용하여 최적의 정책을 찾는다. 이는 MDP 모델을 아래와 같이 큐함수에 대입하여 얻을 수 있다. 아래 식에 따라 모든 상태와 액션에 따른 큐값이 계산된다.

[0039] $$Q(S_\tau, A_\tau) = (1-\alpha)Q(S_\tau, A_\tau) + \alpha(R_{\tau+1} + \gamma \times \max Q(S_{\tau+1}, A_{\tau+1}))$$

[0040] 상기 수학식에서, $0 \le \gamma \le 1$는 디스카운트 팩터(discount factor)를, $0 \le \alpha \le 1$는 학습율(learning rate)을 나타낸다. $\gamma$와 α는 사전이 미리 정해진 값을 가질 수 있다.

[0041] 상기 큐함수는 단위 시간 τ 마다 액션 A를 선택하고, 보상 R을 받으며 새로운 상태 $S_{\tau+1}$로 전이하고, 큐값이 갱신된다. 최적의 정책을 찾은 큐함수는 아래와 같이 나타낼 수 있다.

[0042] $$Q_*(S, A) = \max_\pi [Q_\pi(S, A)]$$

[0043] 상기 수학식에서 π는 정책으로, 각 상태에서 어떤 액션을 취할지 선택하도록 한다. 본 발명에서는 최적의 정책을 얻기 위해 MDP 모델을 큐러닝(Q-Learning)으로 변경하였고, 최적의 정책은 위 문제를 해결함으로써 도출될 수 있다. 위 수학식에서 도출된 결과는 자원 사용을 최소화하고 태스크 요구 지연 시간은 만족하는 최적의 추론 모델 선택이다.

[0044] 본 발명에서 고려하는 환경은 태스크 요청 속도가 동적이고 컴퓨팅 자원은 한정된 모든 네트워크 환경을 가정한다. 이는 모바일의 이동성 및 RAN 자원 사용량 예측뿐만 아니라 네트워크 슬라이스 부하 예측, 컨텐츠 캐싱 등 태스크 발생 속도가 동적이고 추론 가능한 모바일, 서버 등의 네트워크 장치 전반에서 적용 사례를 만들 수 있다. 이는 특히 태스크 요청 속도가 높은 환경에서 큰 효율을 보이고 자원 관리를 용이하게 하여 인프라 관리 및 비용 측면에서 장점을 가진다.

[0045] 도 2와 도 3은 단위 시간 당 태스크 요청 개수가 각각 1, 2일 때 시간 경과에 따른 평균 보상 변화를 보여준다. 본 시뮬레이션에서는 제안 기법에서 태스크가 2개부터 최대 4개까지 MTL 모델로 함께 추론될 수 있도록 하였다. 시뮬레이션은 태스크 요청 속도에 따른 보상 변화에 대해 기존의 추론 모델을 선택하는 과정이 없이 STL 모델만 사용하는 기법(빨간색)과 제안 기법(파란색)을 비교한다. 도면 2, 3의 결과와 같이 두 실험 결과에서 제안 기법이 기존 기법보다 높은 보상을 가지는 것을 확인할 수 있고, 태스크 요청 속도가 높은 경우(도 3)에서 컴퓨팅 자원 한정으로 인해 두 기법 모두 보상이 낮아지는 것을 확인할 수 있다. 이는 제안 기법이 각 태스크의 요구 지연 시간을 넘지 않으면서, 같은 컴퓨팅 자원으로 최대한 많은 태스크를 추론하도록 하는 정책을 적용하였기 때문으로 볼 수 있다.

[0046] 도 4는 도 1에 도시된 추론 모델 선택 장치에 의해 수행되는 추론 모델 선택 방법을 설명하기 위한 흐름도이다. 추론 방법으로 명명될 수도 있는 추론 모델 선택 방법을 설명함에 있어, 앞선 기재와 중복되는 내용에 관하여는 그 구체적인 설명을 생략하기로 한다.

[0047] 우선, 컴퓨팅 자원에 대한 모니터링이 수행된다(S110). 모니터링 동작은 미리 정해진 주기(예컨대, τ초)를 기준으로 주기적으로 수행될 수 있다. 컴퓨팅 파워와 같은 컴퓨팅 자원은 추론 동작에 가용 가능한 컴퓨팅 자원으로써 미리 정해진 규칙에 따라 정량화된 표현으로 도출이 가능하다. 이에 따라, 모델 선택 동작은 주기적으로 업데이트되는 컴퓨팅 자원을 기초로 수행될 수 있다. 추가적으로, 복수의 추론 모델들 각각의 추론을 위한 입력 데이터 역시 주기적으로 또는 비주기적으로 모니터링되거나 외부 디바이스로부터 수신되어 확보되는 것으로 가

정한다.

[0048] 태스크(task)가 수신된다(S120). 태스크는 유무선 통신망을 통하여 외부 디바이스로부터 수신될 수 있다. 이때, 태스크는 해당 태스크의 요구 지연 시간에 대한 정보와 함께 수신될 수 있다. 실시예에 따라, 태스크는 추론 모델의 입력으로 사용되는 데이터를 더 포함할 수도 있다. 또한, 태스크는 임의의 추론 모델에 대한 추론 동작에 관한 요청으로써, 모바일 트래픽 분류, 모바일 트래픽 예측, 모바일 이동성 예측, RAN의 자원 사용량 예측, 네트워크 슬라이스 부하 예측, 컨텐츠 캐싱 등을 포함할 수 있다. 다만, 본 발명의 권리범위가 태스크의 종류에 제한되는 것은 아니며, 태스크는 상술한 태스크 이외에 다양한 태스크일 수 있다. 각 태스크의 요구 지연 시간은 동일하거나 상이할 수 있다.

[0049] 수신되는 태스크를 수행할 추론 모델이 선택된다(S130). 실시예에 따라, 추론 모델의 선택 동작은 미리 정해진 시간 주기, 예컨대 τ 마다 주기적으로 수행될 수 있다. 이 경우, 추론 모델의 선택은 미리 정해진 시간 주기 동안 수신된 태스크에 대한 추론 모델 선택을 의미한다. 각 태스크에 할당되는 추론 모델은 MDP 모델을 큐러닝(Q-Learning)으로 변형한 결과로 도출된 큐함수를 풀이함으로써 도출될 수 있다.

[0050] 추론 모델의 선택이 완료됨에 따라, 각 태스크는 선택된 추론 모델의 큐(queue)에 입력된다(S140). 이때, STL 모델 및/또는 MTL 모델은 복수개가 존재할 수 있고, 이 경우 각 모델에는 대응되는 큐가 구비되어 있다. 따라서, 추론 모델 선택 방법을 수행하는 컴퓨팅 장치에는 복수의 추론 모델들(적어도 하나의 STL 모델과 적어도 하나의 MTL 모델)이 미리 저장되어 있을 수 있다. 또한, 각 추론 모델들은 미리 정해진 추론 동작을 수행할 수 있다.

[0051] 추론 동작이 수행된다(S150). 추론 동작은 추론 모델에 따라 상이하게 수행될 수 있다. 예컨대, STL 모델의 경우, 큐에 입력되는 태스크는 대기 시간 없이 바로 추론 동작이 수행되어, 소정의 추론 시간(컴퓨팅 시간이라 명명될 수도 있음)이 경과한 후에, 추론 결과가 도출될 수 있다. MTL 모델의 경우, 대응하는 큐가 태스크로 가득 찬 경우, 추론 동작이 수행될 수 있다. 다른 예로, MTL 모델의 경우, 큐에 대기하는 태스크 중 요구 지연 시간이 근접한 태스크가 존재한다면, 큐가 태스크로 가득차지 않았음에도 불구하고, 추론 동작이 수행될 수 있다. 일 예로, 태스크가 수신된 시점으로부터 경과된 시간과 추론 모델의 컴퓨팅 시간의 합이 해당 태스크의 요구 지연 시간(또는 요구 지연 시간으로부터 미리 정해진 시간을 뺀 값)과 같아진 경우, 또는 태스크가 큐에 대기한 시간과 추론 모델의 컴퓨팅 시간의 합이 해당 태스크의 요구 지연 시간(또는 요구 지연 시간으로부터 미리 정해진 시간을 뺀 값)과 같아진 경우, 해당 MTL 모델의 추론 동작이 수행될 수 있다.

[0052] 추론 결과는 태스크를 송신한 외부 디바이스로 송신될 수 있다(S160). 각 태스크에 대한 요구 지연 시간을 고려하여 추론 모델을 선택하고, 요구 지연 시간이 경과되기 전에 추론 결과를 도출하였기 때문에, 요구 지연 시간을 만족시키면서 추론 결과를 해당 외부 디바이스로 송신할 수 있다.

[0053] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성 요소는, 예를 들어, 프로세서, 콘트롤러, ALU(Arithmetic Logic Unit), 디지털 신호 프로세서(Digital Signal Processor), 마이크로컴퓨터, FPA(Field Programmable array), PLU(Programmable Logic Unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(Operation System, OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(Processing Element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(Parallel Processor)와 같은, 다른 처리 구성(Processing Configuration)도 가능하다.

[0054] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 구체화(Embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매

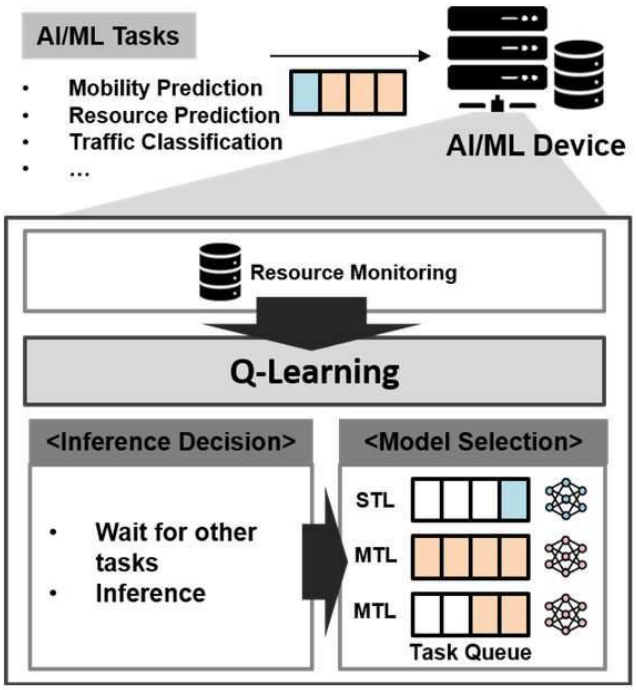
체에 저장될 수 있다.

[0055] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-optical Media), 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

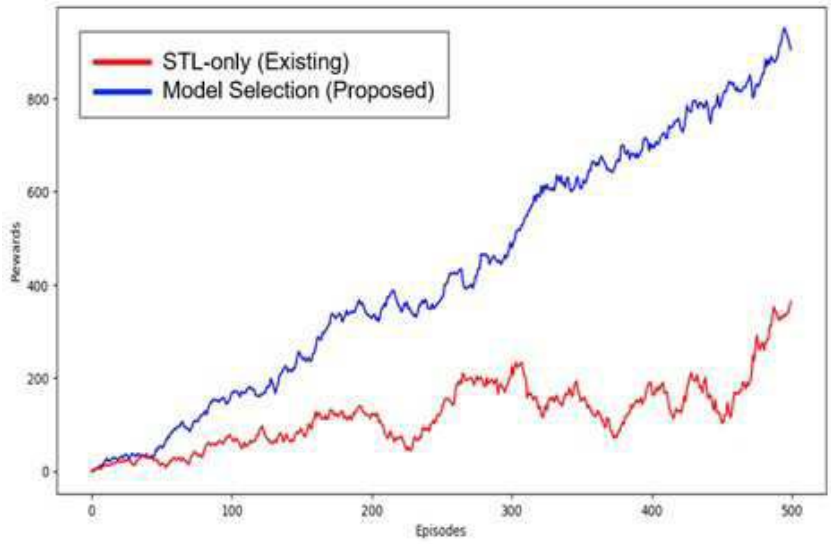
[0056] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

## 도면

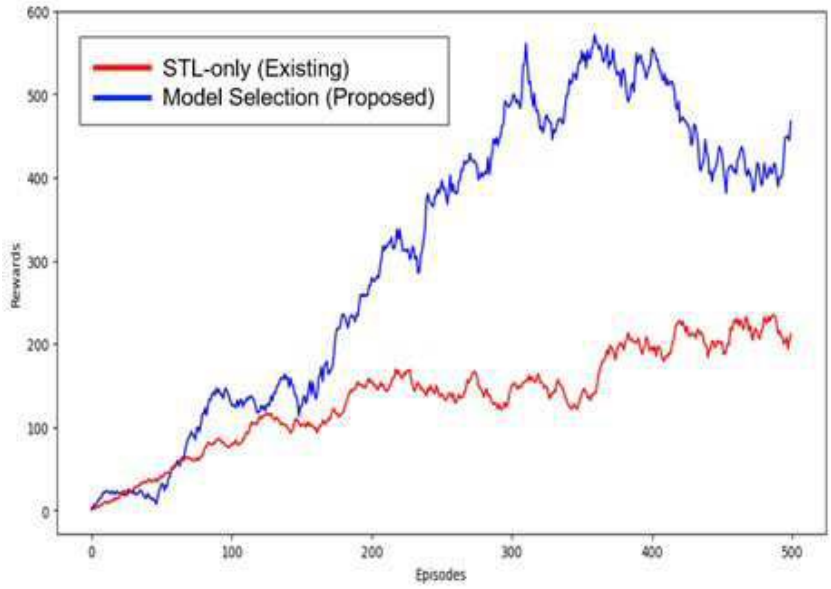
### 도면1



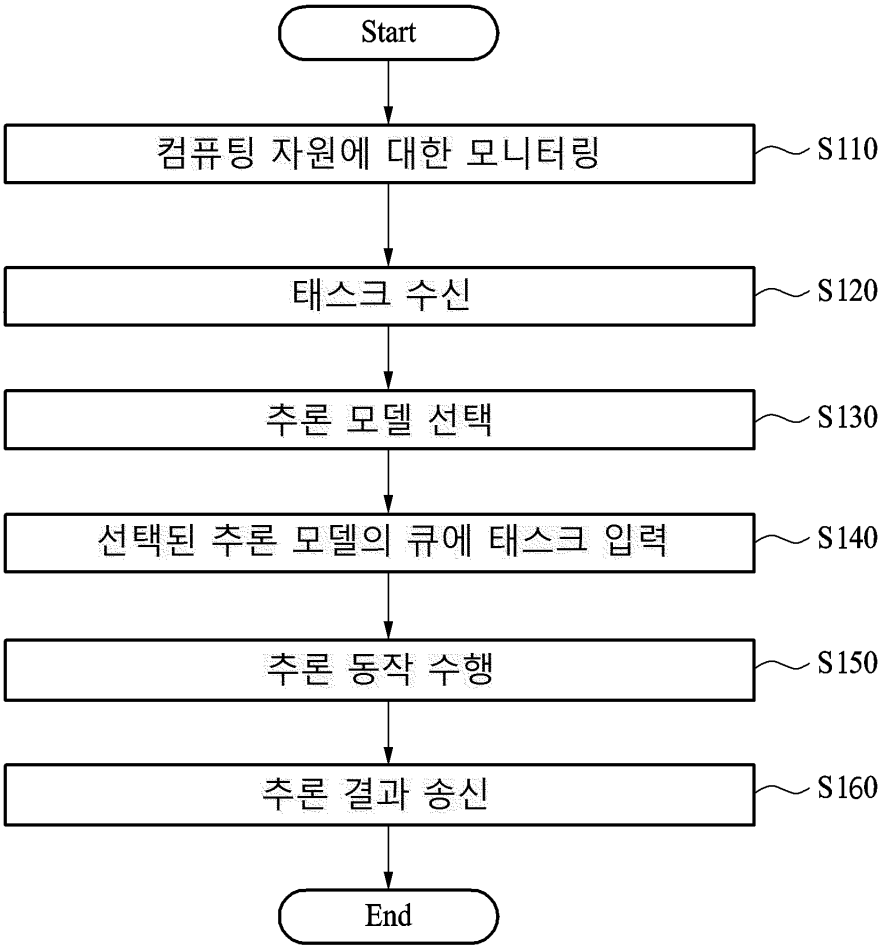
도면2


STL-only (Existing)
Model Selection (Proposed)
Rewards
Episodes
0
200
400
600
800
0
100
200
300
400
500

도면3


STL-only (Existing)
Model Selection (Proposed)
Rewards
Episodes
0
100
200
300
400
500
600
0
100
200
300
400
500

도면4


Start
컴퓨팅 자원에 대한 모니터링
S110
태스크 수신
S120
추론 모델 선택
S130
선택된 추론 모델의 큐에 태스크 입력
S140
추론 동작 수행
S150
추론 결과 송신
S160
End