(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0131142
(43) 공개일자 2025년09월02일

(51) 국제특허분류(Int. Cl.)
H04W 24/02 (2009.01) *G06N 3/096* (2023.01)
H04L 65/40 (2022.01) *H04W 28/02* (2009.01)
H04W 8/24 (2009.01)
(52) CPC특허분류
H04W 24/02 (2013.01)
G06N 3/096 (2023.01)
(21) 출원번호 10-2024-0027504
(22) 출원일자 2024년02월26일
심사청구일자 없음

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
(72) 발명자
유성열
경기도 수원시 영통구 삼성로 129(매탄동)
입에릭
경기도 수원시 영통구 삼성로 129(매탄동)
(74) 대리인
리앤목특허법인

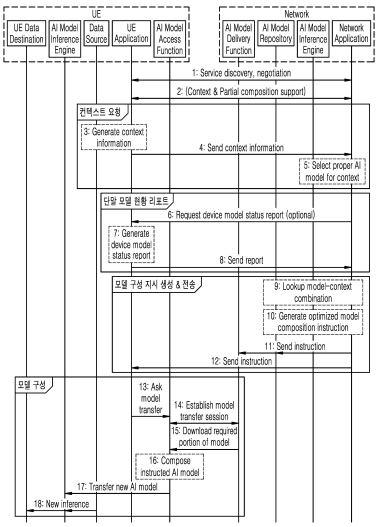
전체 청구항 수 : 총 1 항

(54) 발명의 명칭 **무선 통신 시스템에서 인공지능 모델 업데이트 방법 및 장치**

(57) 요 약

본 개시의 일 실시예에서는, 무선 통신 시스템에서 단말이 동작하는 방법이 제공된다. 방법은, 입력 데이터의 컨텍스트 변경을 식별하는 단계, 서비스 사업자로부터 변경된 컨텍스트의 추론을 위한 인공지능 모델과 관련된 정보를 수신하는 단계, 수신된 정보에 기초하여 인공지능 모델을 식별하는 단계, 및 인공지능 모델을 통해 변경된 컨텍스트의 추론을 수행하는 단계를 포함할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
H04L 65/40 (2022.05)
H04L 67/30 (2022.05)
H04W 28/0215 (2023.01)
H04W 8/24 (2013.01)

명 세 서

청구범위

청구항 1

무선 통신 시스템에서 단말이 동작하는 방법으로서,

입력 데이터의 컨텍스트 변경을 식별하는 단계;

서비스 사업자로부터 변경된 컨텍스트의 추론을 위한 인공지능 모델과 관련된 정보를 수신하는 단계;

상기 수신된 정보에 기초하여 상기 인공지능 모델을 식별하는 단계; 및

상기 인공지능 모델을 통해 상기 변경된 컨텍스트의 추론을 수행하는 단계를 포함하는, 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 무선 통신 시스템에서 인공지능(artificial intelligence, AI) 모델을 업데이트하는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 무선 통신 세대를 거듭하면서 발전한 과정을 돌아보면 음성, 멀티미디어, 데이터 등 주로 인간 대상의 서비스를 위한 기술이 개발되어 왔다. 5G (5th-generation) 통신 시스템 상용화 이후 폭발적인 증가 추세에 있는 커넥티드 기기들이 통신 네트워크에 연결될 것으로 전망되고 있다. 네트워크에 연결된 사물의 예로는 차량, 로봇, 드론, 가전제품, 디스플레이, 각종 인프라에 설치된 스마트 센서, 건설기계, 공장 장비 등이 있을 수 있다. 모바일 기기는 증강현실 안경, 가상현실 헤드셋, 홀로그램 기기 등 다양한 폼팩터로 진화할 것으로 예상된다. 6G (6th-generation) 시대에는 수천억 개의 기기 및 사물을 연결하여 다양한 서비스를 제공하기 위해, 개선된 6G 통신 시스템을 개발하기 위한 노력이 이루어지고 있다. 이러한 이유로, 6G 통신 시스템은 5G 통신 이후 (beyond 5G) 시스템이라 불리어지고 있다.

[0003] 2030년쯤 실현될 것으로 예측되는 6G 통신 시스템에서 최대 전송 속도는 테라 (즉, 1,000기가) bps, 무선 지연시간은 100마이크로초(μsec) 이다. 즉, 5G 통신 시스템대비 6G 통신 시스템에서의 전송 속도는 50배 빨라지고 무선 지연시간은 10분의 1로 줄어든다.

[0004] 이러한 높은 데이터 전송 속도 및 초저(ultra low) 지연시간을 달성하기 위해, 6G 통신 시스템은 테라헤르츠(terahertz) 대역 (예를 들어, 95기가헤르츠(95GHz)에서 3테라헤르츠(3THz)대역과 같은)에서의 구현이 고려되고 있다. 테라헤르츠 대역에서는 5G에서 도입된 밀리미터파(mmWave) 대역에 비해 더 심각한 경로손실 및 대기흡수 현상으로 인해서 신호 도달거리, 즉 커버리지를 보장할 수 있는 기술의 중요성이 더 커질 것으로 예상된다. 커버리지를 보장하기 위한 주요 기술로서 RF(radio frequency) 소자, 안테나, OFDM (orthogonal frequency division multiplexing)보다 커버리지 측면에서 더 우수한 신규 파형(waveform), 빔포밍(beamforming) 및 거대 배열 다중 입출력(massive multiple-input and multiple-output; massive MIMO), 전차원 다중 입출력(full dimensional MIMO; FD-MIMO), 어레이 안테나(array antenna), 대규모 안테나(large scale antenna)와 같은 다중 안테나 전송 기술 등이 개발되어야 한다. 이 외에도 테라헤르츠 대역 신호의 커버리지를 개선하기 위해 메타물질(metamaterial) 기반 렌즈 및 안테나, OAM(orbital angular momentum)을 이용한 고차원 공간 다중화 기술, RIS(reconfigurable intelligent surface) 등 새로운 기술들이 논의되고 있다.

[0005] 또한 주파수 효율 향상 및 시스템 네트워크 개선을 위해, 6G 통신 시스템에서는 상향링크(uplink)와 하향링크(downlink)가 동일 시간에 동일 주파수 자원을 동시에 활용하는 전이중화(full duplex) 기술, 위성(satellite) 및 HAPS(high-altitude platform stations)등을 통합적으로 활용하는 네트워크 기술, 이동 기지국 등을 지원하고 네트워크 운영 최적화 및 자동화 등을 가능하게 하는 네트워크 구조 혁신 기술, 스펙트럼 사용 예측에 기초한 충돌 회피를 통한 동적 주파수 공유 (dynamic spectrum sharing) 기술, AI (artificial intelligence)를 설계 단계에서부터 활용하고 종단간(end-to-end) AI 지원 기능을 내재화하여 시스템 최적화를 실현하는 AI 기반

통신 기술, 단말 연산 능력의 한계를 넘어서는 복잡도의 서비스를 초고성능 통신과 컴퓨팅 자원(mobile edge computing (MEC), 클라우드 등)을 활용하여 실현하는 차세대 분산 컴퓨팅 기술 등의 개발이 이루어지고 있다. 뿐만 아니라 6G 통신 시스템에서 이용될 새로운 프로토콜의 설계, 하드웨어 기반의 보안 환경의 구현 및 데이터의 안전 활용을 위한 메커니즘 개발 및 프라이버시 유지 방법에 관한 기술 개발을 통해 디바이스 간의 연결성을 더 강화하고, 네트워크를 더 최적화하고, 네트워크 엔티티의 소프트웨어화를 촉진하며, 무선 통신의 개방성을 높이려는 시도가 계속되고 있다.

[0006] 이러한 6G 통신 시스템의 연구 및 개발로 인해, 사물 간의 연결뿐만 아니라 사람과사물 간의 연결까지 모두 포함하는 6G 통신 시스템의 초연결성(hyper-connectivity)을 통해 새로운 차원의 초연결 경험(the next hyper-connected experience)이 가능해질 것으로 기대된다. 구체적으로 6G 통신 시스템을 통해 초실감 확장 현실(truly immersive extended reality; truly immersive XR), 고정밀 모바일 홀로그램(high-fidelity mobile hologram), 디지털 복제(digital replica) 등의 서비스 제공이 가능할 것으로 전망된다. 또한 보안 및 신뢰도 증진을 통한 원격 수술(remote surgery), 산업 자동화(industrial automation) 및 비상 응답(emergency response)과 같은 서비스가 6G 통신 시스템을 통해 제공됨으로써 산업, 의료, 자동차, 가전 등 다양한 분야에서 응용될 것이다.

발명의 내용

해결하려는 과제

[0007] 본 개시의 일 실시예는, 무선 통신 시스템에서 서비스를 효과적으로 제공할 수 있는 방법 및 장치를 제공할 수 있다.

과제의 해결 수단

[0008] 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서 개시된, 무선 통신 시스템에서 단말이 동작하는 방법은, 입력 데이터의 컨텍스트 변경을 식별하는 단계, 서비스 사업자로부터 변경된 컨텍스트의 추론을 위한 인공지능 모델과 관련된 정보를 수신하는 단계, 수신된 정보에 기초하여 인공지능 모델을 식별하는 단계, 및 인공지능 모델을 통해 변경된 컨텍스트의 추론을 수행하는 단계를 포함할 수 있다.

도면의 간단한 설명

[0009] 도 1은 본 개시의 일 실시예에 따른 서비스 사업자와 단말의 프로세스를 도시한 흐름도이다.

도 2는 본 개시의 일 실시예에 따른 네트워크 엔티티를 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0010] 본 명세서에서 실시예를 설명함에 있어서 본 발명이 속하는 기술 분야에 익히 알려져 있고 본 발명과 직접적으로 관련이 없는 기술 내용에 대해서는 설명을 생략한다. 이는 불필요한 설명을 생략함으로써 본 발명의 요지를 흐리지 않고 더욱 명확히 전달하기 위함이다.

[0011] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있다.

[0012] 이하 설명에서 사용되는 접속 노드(node)를 식별하기 위한 용어, 망 객체(network entity, 네트워크 엔티티)들을 지칭하는 용어, 메시지들을 지칭하는 용어, 망 객체들 간 인터페이스를 지칭하는 용어, 다양한 식별 정보들을 지칭하는 용어 등은 설명의 편의를 위해 예시된 것이다. 따라서, 본 개시가 후술되는 용어들에 한정되는 것은 아니며, 동등한 기술적 의미를 가지는 대상을 지칭하는 다른 용어가 사용될 수 있다.

[0013] 이하 설명의 편의를 위하여, 본 개시는 3GPP NR(3rd Generation Partnership Project New Radio) 규격에서 정의하고 있는 용어 및 명칭들을 사용한다. 하지만, 본 개시가 상기 용어 및 명칭들에 의해 한정되는 것은 아니며, 다른 규격에 따르는 시스템에도 동일하게 적용될 수 있다. 본 개시에서 기지국은 gNB를 나타낼 수 있다. 또한 단말이라는 용어는 핸드폰, NB-IoT 기기들, 센서들, 뿐만 아니라 또 다른 무선 통신 기기들을 나타낼 수 있다.

[0014] 이하, 기지국은 단말의 자원 할당을 수행하는 주체로서, gNode B, eNode B, Node B, BS(Base Station), 무선

접속 유닛, 기지국 제어기, 또는 네트워크 상의 노드 중 적어도 하나일 수 있다. 단말은 UE(User Equipment), MS(Mobile Station), 셀룰러폰, 스마트폰, 컴퓨터, 또는 통신기능을 수행할 수 있는 멀티미디어시스템을 포함할 수 있다. 물론 상기 예시에 제한되는 것은 아니다.

[0015] 본 개시는 3GPP NR(5세대 이동통신 표준)에 적용할 수 있다. 또한 본 개시는 5G 통신 기술 및 IoT 관련 기술을 기반으로 지능형 서비스(예를 들어, 스마트 홈, 스마트 빌딩, 스마트 시티, 스마트 카 또는 커넥티드 카, 헬스케어, 디지털 교육, 소매업, 보안 및 안전 관련 서비스 등)에 적용될 수 있다.

[0016] 무선 통신 시스템은 초기의 음성 위주의 서비스를 제공하던 것에서 벗어나 예를 들어, 3GPP의 HSPA(High Speed Packet Access), LTE(Long Term Evolution 또는 E-UTRA(Evolved Universal Terrestrial Radio Access)), LTE-Advanced(LTE-A), LTE-Pro, 3GPP2의 HRPD(High Rate Packet Data), UMB(Ultra Mobile Broadband), 및 IEEE의 802.16e 등의 통신 표준과 같이 고속, 고품질의 패킷 데이터 서비스를 제공하는 광대역 무선 통신 시스템으로 발전하고 있다.

[0017] 광대역 무선 통신 시스템의 대표적인 예로, LTE 시스템에서는 하향링크(downlink, DL)에서는 OFDM(Orthogonal Frequency Division Multiplexing) 방식을 채용하고 있고, 상향링크(uplink, UL)에서는 SC-FDMA(Single Carrier Frequency Division Multiple Access) 방식을 채용하고 있다. 상향링크는 단말(user equipment, UE, 또는 mobile station, MS)이 기지국(eNode B 또는 base station, BS)으로 데이터 또는 제어신호를 전송하는 무선링크를 뜻하고, 하향링크는 기지국이 단말로 데이터 또는 제어신호를 전송하는 무선링크를 뜻한다. 상기와 같은 다중 접속 방식은, 각 사용자 별로 데이터 또는 제어정보를 실어 보낼 시간-주파수 자원을 서로 겹치지 않도록, 즉 직교성(Orthogonality)이 성립하도록, 할당 및 운용함으로써 각 사용자의 데이터 또는 제어정보를 구분한다.

[0018] LTE 이후의 향후 통신 시스템으로서, 즉, 5G 통신시스템은 사용자 및 서비스 제공자 등의 다양한 요구 사항을 자유롭게 반영할 수 있어야 하기 때문에 다양한 요구사항을 동시에 만족하는 서비스가 지원되어야 한다. 5G 통신시스템을 위해 고려되는 서비스로는 향상된 모바일 광대역 통신(eMBB; Enhanced Mobile BroadBand), 대규모 기계형 통신(mMTC; massive Machine Type Communication), 초신뢰 저지연 통신(URLLC; Ultra Reliability Low Latency Communication) 등이 있다.

[0019] 또한, 이하에서 LTE, LTE-A, LTE Pro 또는 5G(또는 NR, 차세대 이동 통신) 시스템을 일례로서 본 발명의 실시예를 설명하지만, 유사한 기술적 배경 또는 채널 형태를 갖는 여타의 통신시스템에도 본 발명의 실시예가 적용될 수 있다. 또한, 본 발명의 실시예는 숙련된 기술적 지식을 가진 자의 판단으로써 본 발명의 범위를 크게 벗어나지 아니하는 범위에서 일부 변형을 통해 다른 통신시스템에도 적용될 수 있다.

[0020] 하기에서 본 개시를 설명함에 있어 관련된 공지 기능 또는 구성에 대한 구체적인 설명이 본 개시의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다. 이하 첨부된 도면을 참조하여 본 개시의 실시 예를 설명하기로 한다.

[0021] 인공 지능(artificial intelligence, AI) 및 기계 학습 기술(machine learning, ML)(이하 AI/ML을 AI로 통칭한다)은 이미지 분류, 음성/얼굴 인식과 같은 기존 애플리케이션부터 비디오 품질 향상과 같은 최신 애플리케이션에 이르기까지 미디어 관련 애플리케이션에 도입되어 일반화 되고 있다.

[0022] 이 분야에 대한 연구가 성숙해짐에 따라 더 많은 양의 데이터를 처리해야 하는 고도의 애플리케이션이 고려되고 있으며, 단말 뿐만 아니라 네트워크의 서버와 협력하는 미디어 응용 프로그램을 위한 AI 아키텍처를 5G 시스템에 도입하기 위한 연구가 계속되고 있다. 예를 들어, 이미지 및 비디오 내의 객체 인식, 스트리밍 비디오의 품질 향상, 음성에 대한 자연어 처리를 포함하여, 여러 사례들이 논의되고 있다.

[0023] AI 추론 및 학습을 위한 시나리오로는 사전 훈련된 AI 모델을 네트워크에서 UE로 제공하는 것, UE에서 또는 UE와 네트워크간 분할해서 추론(inference)하는 것들이 포함된다.

[0024] AI 모델의 구조는 레이어로 구성되고, 레이어는 노드들을 포함하고, 노드는 함수 및 가중치 값을 포함한다.

[0025] AI 모델의 식별은 레이어와 노드의 구성으로 식별될 수 있고, 가중치 값은 모델의 훈련에 사용된 데이터와 학습의 회수에 따라 달라질 수 있다.

[0026] 가중치 값을 저장하기 위한 데이터의 크기는 정확도 및 프로세싱 성능에 비례한다. 즉, 데이터의 크기가 커질수록 정확도가 높아지는 반면, 요구되는 프로세싱 성능도 높아진다.

[0027] AI가 적용되는 시나리오를 위한 아키텍처에서는 AI 모델 저장소, AI 모델 제공 기능, AI 모델 액세스 기능, AI 모델 추론 엔진 및 중간 데이터 (intermediate data) 제공 기능과 같은 몇가지 핵심 구성 요소를 결합하여 5G 네트워크를 통해 AI 모델 및 관련 데이터를 효율적이고 효과적으로 제공할 수 있다.

[0028] 5G 네트워크를 통해 AI 서비스 제공자로부터 제공되는 AI 서비스는 서비스 제공자 서버의 애플리케이션과 단말의 애플리케이션간 통신을 기반으로, 어떤 서비스를 제공하고 받을 것인지 결정하는 단계와 해당 서비스에 적합한 AI 모델의 결정 단계, 단말에서 구동하기에 적합한 AI 모델의 버전(또는 Variant)(예를 들어, high bit depth의 고정확도 모델, quantized bit의 저정확도 모델 등)의 결정 단계 등으로 구성될 수 있다.

[0029] 일 실시예에서, 단일한 AI 모델은 모든 범용의 목적에 사용되기에는 적합하지 않을 수 있다. 따라서, 최적의 서비스를 제공받기 위해서는 서비스 제공자로부터 사용 목적에 최적화된 AI 모델을 수신할 필요가 있다.

[0030] 본 개시의 일 실시예에 따르면, 사용자가 원하는 결과에 맞는 최적의 AI 모델을 선택하기 위해 단말과 서버간 통신하는 방법을 제공할 수 있고, 단말과 서버간 교환해야하는 정보를 표준화할 수 있다.

[0031] 일 실시예에서, 변경 대상의 AI 모델과 일부의 차이만이 존재하는 다른 버전의 AI 모델이 단말에 이미 존재할 수도 있고, 변경 대상의 AI 모델 중의 일부만 전송하더라도 단말에 이미 존재하는 다른 AI 모델의 일부와 결합해서 사용 가능한 경우가 있을 수 있다. AI 모델의 선택은 단말에서 임의로 결정할 수 없기 때문에, 보다 효과적으로 전송 동작이 수행되기 위해서 단말의 AI 모델 보유 현황이 서버에 제공되고 서버에 의해 분석될 수 있다.

[0032] 본 개시의 일 실시예에 따른 무선 통신 시스템은, 입력 데이터로부터 데이터 컨텍스트(context)에 따른 추론을 실시할 수 있는 AI 모델(이하, 모델), 다양한 데이터 컨텍스트를 추론할 수 있는 모델들을 구비하고 이를 서비스 가입자에게 제공할 수 있는 서비스 사업자, 및 서비스 사업자로부터 적어도 하나의 모델을 수신하고 이를 사용해 데이터 컨텍스트에 따른 추론을 실시할 수 있는 단말로 구성될 수 있다.

[0033] 일 실시예에서, 단말은 제 1 모델을 사용해 사용자가 추론하고 있던 대상이 변경될 때 직접적으로 대상 또는 컨텍스트를 판단해 서비스 사업자에게 새로운 모델을 요청할 수 있다. 또는, 단말은 취득한 입력 (영상, 음성, 기타 센싱 정보 등)을 서비스 사업자가 요청하는 형태로 전처리한 후 서비스 사업자에게 전달할 수 있고, 서비스 사업자는 적절한 제 2 또는 제 3의 모델을 선정하여 단말에 통지할 수 있다.

[0034] 본 개시의 일 실시예에 따르면, 단말이 AI 모델을 업데이트하는 방법은 제 1 모델을 사용해 사용자가 추론하고 있던 대상이 변경되거나 또는 변화할 때, 변경된 컨텍스트를 판단하는 단계, 변경된 컨텍스트를 추론할 수 있는 적절한 제 2 모델을 선택하는 단계, 및 제 2 모델을 수신하는 단계를 포함할 수 있다. 서비스 제공자가 단말에게 제 2 모델을 제공할 때, 단말이 이미 수신하여 보유하고 있는 제 1 모델 또는 적어도 하나의 제 3 모델의 정보를 기반으로 제 2 모델에 근접한 추론 성능을 제공하는 제 2-호환 모델을 단말에서 구성할 수 있는 방법을 제공하도록 할 수도 있다.

[0035] 각각의 인공지능 모델들은 아래의 용도를 가질 수 있다.

[0036] 제 1 모델: 제 1 컨텍스트 추론 용도

[0037] 제 2 모델: 제 2 컨텍스트 추론 용도

[0038] 제 3 모델: 제 3 컨텍스트 추론 용도

[0039] 제 2-호환 모델: 제 1 모델 또는 제 3 모델의 일부, 및 제 2 모델의 일부에 기초하여 제 2 모델의 추론과 같거나 근접한 추론 성능을 제공하는 용도

[0040] 본 개시에서 컨텍스트(context)란, 사용자가 의도하고 수신하고자 하는 정보들간의 연관성 및 맥락이 유지되는 단위를 나타낼 수 있다. 예를 들어, 꽃의 종류가 일 컨텍스트라 하면 다양한 꽃의 이미지들이 하나의 컨텍스트가 되고 자동차의 종류가 또 다른 일 컨텍스트라 하면 다양한 자동차의 이미지들이 하나의 컨텍스트가 될 수 있다. 예를 들어, 꽃의 종류를 판별하는 모델은 자동차의 종류를 판별하기 위해 사용되기에 적절하지 않기 때문에, 추론하고자 하는 대상이 변경되는 시점에 변경된 대상의 컨텍스트를 판별할 수 있는 모델로 단말이 이용하는 인공지능 모델이 변경될 수 있다.

[0041] 컨텍스트의 변경은 사용자에 의해 의도될 수도 있고, 추론하려는 대상에 대해 현재 이용중인 인공지능 모델이 적합한 추론 결과를 찾지 못할 때 컨텍스트의 변경이 있는 것으로 식별될 수도 있다. 사용자가 컨텍스트를 변경하는 경우, 변경된 컨텍스트가 서비스 사업자에 명시적으로 요청될 수 있으며, 서비스 사업자는 변경된 컨텍스

트를 추론할 수 있는 가장 적절한 모델을 내부 저장소에서 찾아 단말에게 제공할 수 있다. 컨텍스트가 서비스 사업자에 의해 판단되고 변경되는 경우, 단말에서 추론을 위해 취득한 입력 데이터의 전체 또는 일부가 서비스 제공자에게 전달될 수 있다. 입력 데이터는 압축되거나 특징점을 추출하는 등으로 전처리될 수 있다. 서비스 사업자는 전처리된 입력 데이터를 수신한 후 복호할 수 있고, 복호한 정보로부터 입력 데이터가 가지고 있는 컨텍스트를 판단할 수 있다. 단말은 서비스 사업자가 요구하는 형식으로 입력 데이터를 전처리 할 수 있는 지 판단할 수 있고, 서비스 사업자가 요구하는 복수 개의 전처리 방법 중 지원할 수 있는 적어도 한 가지 형식을 서비스 사업자에게 보고할 수 있다. 서비스 사업자는 단말이 전처리하여 전달하는 입력 데이터에 대한 복호화를 수행하고, 입력 데이터의 컨텍스트를 판단하고, 컨텍스트가 변경되었다고 판단되는 경우 변경된 컨텍스트 판단에 적절한 제 2 모델을 선택할 수 있다. 서비스 사업자는 가능한 경우 단말이 기 보유하고 있는 모델들로부터 제 2 모델과 같거나 근접한 모델을 구성하도록 하는 방법을 지시할 수 있다. 서비스 사업자는 선택된 제 2 모델을 단말에 전송하여 단말이 사용하도록 지시할 수도 있다. 단말은 서비스 사업자로부터 제 2 모델의 전체를 수신하거나, 또는 제 2 모델의 일부와 제 2-호환 모델을 구성하기 위한 정보를 수신할 수 있다.

[0042] 일 실시예에서, 단말은 단말의 모델 수신부와 네트워크의 모델 전송부간 전송 세션을 체결한 후 모델을 수신할 수 있다. 단말은 수신한 전체 또는 일부의 모델로부터 변경된 컨텍스트에 대한 추론을 실시할 수 있는 제 2 모델을 식별하거나 또는 제 2-호환 모델을 구성할 수 있다. 단말은 식별되거나 구성된 모델을 사용해 추론 엔진을 통해 입력 데이터에 대한 추론을 실시할 수 있다.

[0043] 본 개시의 일 실시예에서 인공지능 모델의 구성이란, 인공지능 모델을 구성하는 계층(layer)들을 재위치시키고 다른 인공지능 모델의 적어도 하나의 계층과 결합하여 새로운 모델을 생성하는 동작을 나타낼 수 있다.

[0044] 예를 들어, 단말은 일상의 물체를 식별하는 범용적인 용도의 100개의 계층을 가진 제 1 모델을 구비하고 있을 수 있다. 100개의 계층을 가진 또 다른 제 2 모델이 특정한 카테고리에 한정된 추론에 이용되는 경우, 제 1 모델의 일부 계층(예를 들어, 1-90계층)과 제 2 모델의 일부 계층(예를 들어, 91-100계층)을 결합해 생성한 제 2-호환 모델(100개 계층)을 구성하고 이를 사전에 학습해둔 후, 단말에서 제 2 모델을 필요로 할 때, 단말에서의 제 2-호환 모델의 구성에 이용될 10개 계층만을 전송하고 단말이 이미 구비하고 있던 제 1 모델과 결합하도록 지시함으로써, 단말에서 제 2-호환 모델을 복원하도록 할 수 있다.

[0045] 도 1은 본 개시의 일 실시예에 따른 서비스 사업자(네트워크)와 단말의 프로세스를 도시한 흐름도이다.

[0046] 단계 1에서, 단말 애플리케이션과 네트워크 애플리케이션간 통신을 통해 서비스 사업자가 제공하는 서비스의 개요가 전달되고 사용자에 의해 서비스가 선택될 수 있다.

[0047] 단계 2에서, 단말과 네트워크는 컨텍스트의 교환과 인공지능 모델의 구성이 지원되는지 상호 판단할 수 있다.

[0048] 단계 3에서, 단말과 네트워크 간의 컨텍스트의 교환이 가능한 경우, 단말 애플리케이션은 입력 데이터로부터 컨텍스트 정보를 생성할 수 있다. 예를 들어, 단말은 입력 데이터로부터 컨텍스트의 전처리를 수행할 수 있다.

[0049] 단계 4에서, 생성된 컨텍스트 정보는 네트워크로 전달될 수 있다.

[0050] 단계 5에서, 네트워크 애플리케이션은 전처리되어 있던 컨텍스트 정보를 복원하고, 네트워크의 추론 엔진을 사용해 입력 데이터가 포함하고 있는 컨텍스트를 판단하고, 해당 컨텍스트를 추론할 수 있는 적절한 인공지능 모델을 선택할 수 있다.

[0051] 단계 6에서, 네트워크 애플리케이션은 단말에게 단말이 기 보유하고 있는 인공지능 모델의 정보를 요청할 수 있다. 일 실시예서, 단말은 네트워크의 요청이 없어도 필요한 경우 인공지능 모델 정보를 네트워크에게 보고할 수 있다.

[0052] 단계 7에서, 단말은 보유하고 있는 인공지능 모델들의 정보를 수집해 리포트를 생성할 수 있다.

[0053] 단계 8에서, 단말은 보유하고 있는 모델들의 정보를 네트워크 애플리케이션으로 보고할 수 있다.

[0054] 단계 9에서, 네트워크 애플리케이션은 모델 저장소를 조회하여 컨텍스트를 추론할 수 있는 모델들의 목록을 수집할 수 있다.

[0055] 단계 10에서, 네트워크 애플리케이션과 모델 저장소는 단말이 보유하고 있는 모델 목록을 기반으로 컨텍스트 추론에 최적의 인공지능 모델 또는 호환 모델을 구성하기 위한 방법의 지시를 생성할 수 있다.

[0056] 단계 11에서, 생성된 지시가 네트워크 애플리케이션으로부터 모델 저장소에 전달되고, 모델 저장소에서 모델 전

송부로 전달될 수 있다.

[0057] 단계 12에서, 네트워크 애플리케이션은 단말 애플리케이션에 지시 방법을 전달할 수 있다.

[0058] 단계 13에서, 단말 애플리케이션은 인공지능 모델의 수신을 위해 모델 수신부에 모델 전송 세션의 체결을 지시할 수 있다.

[0059] 단계 14에서, 단말의 모델 수신부는 네트워크의 모델 전송부와 통신해 모델 전송 세션을 체결할 수 있다.

[0060] 단계 15에서, 모델 수신부는 수신한 지시에 명시된 모델의 적어도 일부분을 수신할 수 있다.

[0061] 단계 16에서, 모델 수신부는 수신한 지시에 따라, 보유하고 있는 모델과 수신한 모델의 적어도 일부를 결합해 새로운 모델을 생성할 수 있다.

[0062] 단계 17에서, 모델 수신부는 단말의 추론 엔진에 생성된 새로운 모델을 전달할 수 있다.

[0063] 단계 18에서, 입력 데이터로부터 새로운 추론이 수행될 수 있다.

[0064] 이하, 본 개시의 일 실시예에 따른 단말에서 입력 데이터의 컨텍스트에 맞게 AI 모델을 교체하여 추론에 사용하는 방법과 관련해 다음을 각각 구체적으로 설명한다.

[0065] - 컨텍스트 요청 방법

[0066] - 단말 보고 정보

[0067] - 서버 판단 절차, 조합 모델 구성 지시 방법

[0068] **컨텍스트 요청 방법**

[0069] 일 실시예에서, 서비스 사업자(네트워크, 서버)와 단말 간 교환되는 컨텍스트 요청을 위해 메시지가 사용될 수 있다. 메시지에는 서비스 사업자가 지원하는 컨텍스트 형식 체계들과 그 하위에 포함되는 컨텍스트 형식 식별자 등을 포함할 수 있다. 컨텍스트 형식 체계는 이를 식별하기위한 컨텍스트 형식 체계 식별자로써 식별될 수 있다. 예컨대 컨텍스트 형식 체계 식별자가 urn:3gpp:ai4media:context:cset=15 인 경우, 컨텍스트 형식 식별자는 3GPP의 ai4media 규격에서 지원하는 컨텍스트로 정의한 컨텍스트 형식 체계 식별자 중 하나이고 컨텍스트 모음(context set)의 버전은 15라고 이해될 수 있다. 더 높은 숫자를 사용해 표시하는 상위 버전의 컨텍스트 모음은 낮은 숫자의 하위 버전의 컨텍스트 모음에 비해 추가적인 컨텍스트가 포함되었음을 의미한다. 만약 단말이 하위 버전의 컨텍스트 모음에 기반해 컨텍스트를 조회하고 요청했다면 서비스 사업자는 상위 버전의 컨텍스트 모음을 제안할 수 있고, 단말은 상위 버전의 컨텍스트 체계를 수신한 이후 새로운 컨텍스트를 조회하고 이를 요청할 수 있다.

[0070] 서비스 애플리케이션은 서비스 개시 시점 또는 단말의 필요에 의한 요구에 따라 서비스 애플리케이션에서 지원할 수 있는 컨텍스트 형식 체계 식별자의 목록을 제공할 수 있다.

[0071] 단말 애플리케이션은 서비스 애플리케이션으로부터 수신한 컨텍스트 형식 체계 식별자로부터 가용한 컨텍스트 형식 식별자들의 목록을 수신하고 목록에 포함된 컨텍스트들 중 각각 또는 여러 개를 지원하는 모델의 전송을 서비스 애플리케이션에 요청할 수 있다.

[0072] 단말 및 단말을 사용하는 사용자가, 원하는 컨텍스트를 조회할 수 있도록 컨텍스트 형식 식별자는 사람이 읽을 수 있는 텍스트 형태의 컨텍스트 형식 서술자를 제공할 수 있다. 예를 들어, 사용자는 찾고자 하는 컨텍스트를 서술자로부터 읽어 선택하고, 단말 애플리케이션은 선택된 서술자에 해당하는 컨텍스트 형식 식별자를 서비스 애플리케이션에 요청할 수 있다.

[0073] 컨텍스트 요청 메시지의 구성 요소는 아래 표 1과 같이 나타낼 수 있다:

표 1

[0074]

이름	의미
컨텍스트 형식 체계 식별자	URN등 선언된 컨텍스트의 종류 체계(scheme) 식별자. (예: urn:3gpp:ai4media:context:cset=15)
컨텍스트 형식 서술자	사람이 읽을 수 있는 텍스트 형태의 컨텍스트 식별자 (예: 꽃, Flower)
컨텍스트 형식 식별자	명시적으로 지시 및 요청하는 컨텍스트 (예: 꽃, 자동차, 표정)

컨텍스트 프로파일 서술자	사람이 읽을 수 있는 텍스트 형태의 컨텍스트 프로파일 (예: 한국의 꽃들, Korean flowers)
컨텍스트 프로파일 식별자	컨텍스트 내 종류 또는 범위 (예: 한국의 꽃, 전세계의 꽃)

[0075] 3GPP TR 26.927의 Service requirement information에는 서비스 사업자가 제공하는 서비스의 형태에 관한 정보가 포함된다. 컨텍스트에 관련된 정보는 아래의 표 2와 같이 나타낼 수 있다.

표 2

[0076]

Metadata category	Metadata type	Definition	Metadata type description (Examples)
Service requirement information	Maximum service inference latency	The maximum inference latency requirement specified for a given AI media service, in milliseconds. In the case of split inferencing, this requirement includes the delivery latency of the intermediate data between the first and second split inference entities.	100ms
	Minimum service inference accuracy	The minimum accuracy specified for a given AI media service.	80%
	Service type identifier	An identifier for the service type to be supported by the AI/ML model, such as ASR (Automatic Speech Recognition), TTS (Text To Speech), Translation (with the indication of input and output languages).	TTS, ASR, Trans-EN-to-ZH
	Context type scheme identifier (컨텍스트 형식 체계 식별자)	An identifier specifying a list of supportable contexts by the AI/ML model. A URN is used as the identifier of the list and the corresponding list is managed by service provider.	URN
	Context Scheme	structured information of context scheme. It may include multiple context_types where each context_type has multiple context_profiles.	structured text
	Service accuracy	The expected service accuracy	85%

[0077] 표 2: Service requirement information

[0078] 구조 형태로 표현한 컨텍스트 체계의 구성은 아래 표 3과 같이 나타낼 수 있다. Context_scheme에는 여러 개의 context_type (identifier 및 descriptor)가 포함될 수 있고, context_type 아래에는 여러 개의 context_profile (identifier 및 descriptor)가 포함될 수 있다.

표 3

[0079]

```
metadata:context_scheme
   context_scheme_idenfier: string
       context_type{
           context_type_idenfier: value
           context_type_descriptor: string
           context_profile{
               context_profile_identifier: value
               context_profile_descriptor: string
           }
       }
```

[0080] 일 실시예에서, 입력 데이터의 컨텍스트가 사용자에 의해 명시적으로 지시되지 않고, 서비스 제공자에 의해 판단되어질 수도 있다. 이를 위해, 서비스 제공자는 단말 애플리케이션으로 하여금 전체 또는 일부의 입력 데이터

에 대해 그대로 또는 전처리하여 전송하도록 지시할 수 있다. 서비스 애플리케이션은 서버 내에 포함된 추론 엔진을 사용하여, 단말로부터 수신한 입력 데이터가 포함하고 있는 컨텍스트를 유추할 수 있다.

[0081] 서비스 애플리케이션은 단말 애플리케이션이 실시해야 하는 전처리의 종류를 선택하고 이를 지시할 수 있다. 컨텍스트 전처리 체계 식별자는 단말에서 입력 데이터에 대해 실시가 요청되는 컨텍스트 전처리 방법들이 포함된 목록 체계의 식별자를 나타낼 수 있고, 컨텍스트 전처리 방법은 목록 체계에 속하는 하나의 방법일 수 있다. 전처리는 전체 또는 일부의 입력 데이터에 대해 한정적으로 실시될 수 있으며, 이 때, 하나 이상의 전처리를 나타내는 식별자와 해당 식별자를 서술하는 속성 값이 단말에 제공될 수 있다.

[0082] 단말 애플리케이션은 컨텍스트 전처리 체계 식별자를 서비스 애플리케이션과 교섭하는 과정에서 체계 내 포함된 컨텍스트들 중에서 지원 가능한 컨텍스트 전처리 체계 식별자를 선택해 서비스 애플리케이션에 보고하거나, 또는 지원 가능한 컨텍스트 전처리 체계 식별자가 없음을 보고할 수 있다. 단말이 지원 가능한 컨텍스트 전처리 방법이 없다고 서비스 애플리케이션에 보고한 경우, 서비스 애플리케이션은 단말에 다른 컨텍스트 전처리 체계 식별자를 제시할 수 있다.

[0083] 교섭 이후 단말 애플리케이션은 입력 데이터에 대해 전처리를 실시할 수 있다. 전처리는 전처리 한정 식별자 및 속성 값에 따라 입력 데이터의 전체 또는 일부에 대해 실시될 수 있다. 단말에서 서버로 전송되는 전처리된 입력 데이터에는 전처리 종류와 각 입력 데이터의 처리 단위에 해당하는 미디어 시간, 전처리 개시 및 소요 시간 등이 포함될 수 있다.

[0084] 단말 애플리케이션은 서비스 애플리케이션에 전처리된 컨텍스트 정보, 전처리 정보, 및 전처리 정보의 전송 시간을 전송할 수 있다. 서비스 애플리케이션은 수신된 컨텍스트 정보를 네트워크의 추론 엔진에 전달해 컨텍스트를 추론하고 결정할 수 있다. 네트워크의 추론 엔진은 컨텍스트 정보로부터 컨텍스트를 도출해 서비스 애플리케이션에 회신할 수 있다.

[0085] 서비스 애플리케이션은 입력 데이터 미디어 시간 별로 컨텍스트를 판단해 하나 이상의 인공지능 모델을 선정할 수 있다. 서비스 애플리케이션은 전처리 개시 시간들 간의 간격 등을 고려하여 단말 애플리케이션에서 처리할 전처리 정보의 초당 전송 회수나 해상도, 전처리 파일 크기 등을 조정할 수 있다. 서비스 애플리케이션은 단말이 전처리 정보를 전송한 전송 시간과 네트워크에서 전처리 정보를 수신한 수신 시간과 전처리 정보의 크기를 사용해 전처리 전송 시간을 판단할 수 있으며, 네트워크 리소스의 관리, 단말에 할당된 QoS 등의 요인을 고려해 단말의 초당 전송량, 초당 전송 회수 등을 조정할 수 있다. 서비스 애플리케이션으로부터 변경된 전처리 정책이 수신되면 단말은 변경된 정책에 따라 전처리 종류, 실시 회수 등을 변경할 수 있다.

[0086] 서비스 애플리케이션은 단말에서 보고한 전처리된 컨텍스트 정보 내 서로 다른 미디어 시간에 대해 서로 다른 인공지능 모델을 제시할 수 있다. 즉, 제 1 미디어 시간에 대해 제 1 컨텍스트로 판단하여 제 1 모델을 제시하고, 제 2 미디어 시간에 대해 제 2 컨텍스트로 판단하여 제 2 모델을 제시할 수 있다. 서비스 애플리케이션의 응답은 추론 마다, 미디어 시간 단위마다, 또는 일정 시간 구간 범위 내의 여러 추론들을 묶어서 전달될 수 있다. 따라서, 단말에서 복수 개의 미디어 시간들에 대해 컨텍스트 정보를 생성해 전송할 때, 서비스 애플리케이션은 복수 개의 인공지능 모델들을 사용하도록 응답하고, 이 때 각 미디어 시간별로 서로 다른 하나 이상의 모델을 명시할 수 있다.

[0087] 컨텍스트의 전처리와 관련된 메시지의 구성 요소는 아래 표 4와 같이 나타낼 수 있다:

표 4

[0088]

이름	의미
컨텍스트 전처리 체계 식별자	단말에서 컨텍스트 전처리에 사용할 수 있는 전처리 방법을 정의하는 체계 (예: urn:3gpp:ai4media:context-preprocessings)
컨텍스트 전처리 방법 식별자	단말에서 입력 데이터를 전처리하는 방법의 식별자. (예: 영상 압축, 특징점 추출, 외곽선 추출, 텍스처 추출, embedding, 특정 AI 모델의 초반 n개 계층 실행 등)
전처리 한정 식별자 (Context preprocessing control parameter)	단말에서 전처리 시 한정하는 항목. (예: 해상도, 전처리 파일 크기, 초당 전송 회수, 초당 전송량, 색 공간 범위(RGB, Luma 밝기))
입력 데이터 미디어 timestamp	입력 데이터의 미디어 시간을 나타낸다.
전처리 개시 timestamp	단말에서 전처리가 시작된 시간을 나타낸다.
전처리 전송 timestamp	단말에서 전처리 후 컨텍스트 정보를 전송 시작한 시간을 나타낸다.

[0089] 3GPP TR 26.927의 Endpoint capability information에는 단말 또는 네트워크의 추론 엔진이 수행할 수 있는 기능적 및 성능적 요구사항 정보가 포함될 수 있다. 컨텍스트의 전처리에 관련된 정보는 아래 표 5와 같이 나타낼 수 있다.

표 5

[0090]

Metadata category	Metadata type	Definition	Metadata type description (Examples)
Endpoint capability information	Processing capabilities	The available resources for processing AI/ML model including the computational power (in FLOPS), the memory to store model parameters and perform the inference.	NPU 10TFLOPS, MEM 10GB
	Supported AI Framework	The AI framework(s) supported by the endpoint.	TensorFlow 2.0
	Supported compression algorithms	The supported compression algorithm(s) for intermediate data compression.	NONE, FC_VCM, SNAPPY, ...
	Context preprocessing scheme identifier	An identifier specifying a list of supportable context preprocessing methods by UE or Network. A URN is used as the identifier of the list and the corresponding list is managed by service provider.	URN
	Context preprocessing method identifier	A list of identifiers on preprocessing methods those are supported.	Video compression, feature extraction, edge detection, embedding, split inferencing of specific model
	Context preprocessing control parameter	Preprocessing parameter to control the size, frequency and so on.	512 by 512 resolution
	Connection capabilities	This indicates the available bandwidth in bit/s between the UE and the network for transmitting the AI model and/or the intermediate data.	256 kb/s

[0091] 표 5: Endpoint capability information

[0092] 전처리된 데이터는 intermediate data로서 전송될 수 있으며 전처리된 입력 미디어에 관련된 정보가 포함될 수 있다.

표 6

[0093]

Metadata category	Metadata type	Definition	Metadata type description (Examples)
Intermediate data information	Tensor structure information	The exact underlying tensor structure of the intermediate data tensors including the exact version of it.	PyTorch 2.0, Tensor flow v2.13.0, NumPy v1.25
	Tensor shape	The tensor shape(s) when the output is intermediate data. Tensor shape is a tuple of positive integers, where the size of the tuple represents the dimension of the tensor, and each value represents the size in each dimension.	[1,64,64,64].
	Tensor element data type	The data type of each output intermediate data tensor	:int64, Float32
	Data direction	This defines the direction of transmitted data, either uplink (from UE endpoint to network endpoint) or downlink (From a network endpoint to the UE endpoint). This information may be useful to configure an intermediate data delivery session	Upstream, Downstream
	Compression algorithm	Identifies the compression algorithm(s) that can be applied to the intermediate data. When the connectivity condition between the UE and the network is insufficient to transmit the original intermediate data, a compression algorithm may be applied.	NONE, FC_VCM, SNAPPY, ...
	Input data media timestamp	The media playback time of input data. Server may suggest multiple different contexts per each media timestamp. UE may select one of them, or use different AI models per frame. This timestamp can be relative to media encoder or media playback timeline.	timestamp
	Preprocessing start at timestamp (T1)	The timestamp when preprocessing of input data is started at. Multiple T1 can be included as many as the number of media frames. UE and server may refer same clock to generate the timestamp.	timestamp
	Preprocesing sent at timestamp (T1')	The timestamp when preprocessed data is sent to server. The server may calculate the time difference between sending and receiving to suggest smaller or larger preprocessing result to be sent. UE and server may refer same clock to generate the timestamp.	timestamp

[0094] 표 6: Intermediate data information for split AI/ML operations

[0095] **단말 보고 정보**

[0096] 일 실시예에서, 서비스 애플리케이션은 추론 분석된 컨텍스트 혹은 단말로부터 명시적으로 요청된 컨텍스트에 대해 최적의 인공지능 모델을 선정함에 있어 단말에서 이미 보유하고 있는 인공지능 모델들의 목록을 검토할 수 있다. 서비스 애플리케이션이 이를 위해 단말 애플리케이션에 모델 목록의 전송을 요청하거나, 또는 단말 애플리케이션이 자발적으로 서비스 애플리케이션으로 모델 목록을 전송할 수 있다.

[0097] 모델 목록에는 인공지능 모델 식별자, 인공지능 모델 버전 등의 정보가 포함될 수 있다. 보다 상세하게, 인공지능 모델은 모델이 가지고 있는 각 계층에 대해 식별자를 제공할 수 있고 해당 계층의 버전을 제공할 수 있다. 서비스 제공자는 같은 모델에 대한 추가 학습을 통해 모델을 계층별로 업데이트 할 수 있고, 이 때 특정 계층에 대해서는 업데이트를 억제할 수도 있다. 예컨대 제 1, 2, 3, 4 계층에 대해서는 업데이트가 되지 않도록 값을

고정하고, 제 5 계층 이후에 대해서는 학습에 따라 값을 업데이트할 수도 있다. 이 때, 모델의 계층별로 식별자 및 버전이 고정되거나 변경될 수 있다.

[0098] 서비스 애플리케이션은 단말에서 보유하고 있는 모델의 정보를 수신한 이후, 모델의 적합성, 즉 컨텍스트에 대한 추론의 정확도를 판단할 수 있고, 네트워크의 모델 저장소에서 같은 모델에 대한 새로운 버전이 있는지 판단할 수 있고, 그에 따라 단말에 저장 되어있던 모델에 대한 부분 계층 업데이트를 지시할 수 있다.

[0099] 또한 서비스 애플리케이션은 단말에서 보유하고 있는 모델의 정보를 수신한 이후, 새로운 컨텍스트에 대한 추론을 지원하기 위해서 단말에서 보유하고 있는 모델들 중 활용할 수 있는 모델이 존재하는지, 존재 시 모델의 어떤 계층과, 네트워크의 모델 저장소에서 보유하고 있는 어떤 모델의 어떤 계층을 결합하도록 구성하면 최소한의 데이터 전송을 통해 새로운 컨텍스트에 대한 추론을 지원할 수 있을 지 판단할 수 있다.

[0100] 일 실시예에서, 단말에 저장된 모델과 관련되어 단말에서 전송할 수 있는 정보 구성 요소는 아래 표 7과 같이 나타낼 수 있다:

표 7

[0101]

이름	의미
모델 식별자	모델을 구분하는 식별자 (예: VGG, mobileNet 등)
모델 버전	같은 모델에 대한 서로 다른 버전 정보 (예: 모바일 버전, quantization 단계(float/ 32/ 16/ 8비트) 버전, operation set 버전, 크기 등)
모델 그래프	모델의 입력, 출력, 중간 노드, operation과 그 파라미터 규격에 대한 서술
모델 계층 식별자	모델의 계층별 식별자 (예: layer-01)
모델 계층 버전	모델의 계층별 버전 (예: layer-01-16bit-v01), hash code 등
모델 계층 URL	모델의 특정 계층만을 수신하기 위한 경로 정보

[0102] 3GPP TR 26.927의 AI model information for split AI/ML operations에는 인공지능 모델의 계층별 식별을 위한 정보가 포함된다. 모델 그래프 및 계층 관련 정보는 아래 표 8과 같이 나타낼 수 있다.

표 8

[0103]

Metadata category	Metadata type	Definition	Metadata type description (Examples)
Split model information	Split points	The number of predefined split points at which a certain model can be divided into two for split inferencing.	2
	Model graph	The graph of model which provides outline of the model.	ONNX graph
	Layer identifier	An identifier of the layer in a description of a graph. UE may distinguish whether the same operation and same parameter will be processed as the service provider intended.	layer-01
	Layer version	A version information of a layer. The version number changes with training of new input. UE may distinguish whether the same result will be calculated as the service provider intended.	16bit-v1-build3322, hash code
	Layer URL	URL to download specific layer from network	URL

Split point information	Split point identifier	An identifier of the split point in a description of a computing graph, may be generated by a neural network description language such as ONNX/NNEF. Identifiers must guarantee unique identification of a specific split point.	Nb:10, 75 Name: Layer_10,
	Split point intermediate data size	The size of the intermediate data resulting from the give split point, in kilobytes. Intermediate data size is typically dependent on the tensor size at the given split point.	1086KB
	Split point number	The number of the split point where the split occurs. The number may belong to set of identified numbers defined at the configuration stage.	10
	Split point name	The name of the split point where the split occurs. The name may belong to set of identified split point names defined at the configuration stage.	conv2d_1234
	Split point flag	An information on whether to consider the split point before the split point identifier or after. The convention on whether it is before or after may be defined at the configuration stage.	before, after

[0104] 표 8: AI model information for split AI/ML operations

[0105] **서버 판단 절차, 조합 모델 구성 지시 방법**

[0106] 서비스 애플리케이션은 서비스에서 제공하는 주요 모델들을 기반으로 각 모델의 사용 빈도 및 재활용 가능성 등을 지표화 하고, 이에 따라 두 개 이상의 모델을 결합한 새로운 모델을 학습할 수 있다. 사용자는 컨텍스트를 수시로 변경하거나 복수 개 요청할 수 있으나, 단말에 해당 컨텍스트를 추론할 수 있는 모델이 없는 경우, 추론할 수 있는 모델이 없다고 판단된 이후에 다운로드가 비로소 개시되어 서비스 지연으로 인한 불편이 발생하게 된다. 또한, 인공지능 모델은 수백MB 정도의 크기를 가지므로 일회성의 서비스 제공을 위해 다운로드하고 이용 후 삭제하기에 비효율적이다.

[0107] 인공지능 모델들은 공통의 이미지들 (예를 들어, ImageNet dataset)로 학습되어 있어, 새로운 컨텍스트에 대해 학습하기 위해서는 최종 계층이 해당 새로운 컨텍스트에 대해 미세 조정(fine tuning) 또는 전이 학습(transfer learning)될 수 있다.

[0108] 본 개시의 일 실시예에 따른 서비스 제공자 또는 서비스 애플리케이션은 제공하고자 하는 컨텍스트에 대해 최적의 모델을 제안하기 전에, 단말이 기 보유하고 있는 모델을 확인하고, 기 보유한 모델을 활용하는 방법을 먼저 판단할 수 있다. 즉, 서비스 제공자 또는 서비스 애플리케이션은 단말이 기 보유한 모델의 활용 방법을 단말에게 제공하거나, 또는 신규 모델을 단말에게 제공할 수 있다.

[0109] 서비스 애플리케이션은 제공하고자 하는 컨텍스트를 판단할 수 있고, 서비스에서 제공할 하나 이상의 주요 모델들을 선정할 수 있고, 해당 주요 모델들에 대해 미세 조정 또는 전이 학습 계층을 선정하고 조정/전이를 수행할 수 있다. 이후, 서비스 애플리케이션은 대응되는 모델에 대한 정보와 조정/전이 계층의 조합 정보, 및 조합 모델의 속성 정보 (컨텍스트, 프로세싱 시간, 정확도, 크기)를 식별할 수 있다.

[0110] 식별된 모델의 정보, 모델들의 조합 정보, 및 조합 모델의 속성 정보는 네트워크 또는 서비스 사업자 서버의 저장소 등에 구비되고 서비스 애플리케이션의 요청에 따라 조회되고 제공될 수 있다.

[0111] 조합 모델 정보에는, 모델의 속성인 모델 식별자, 지원하는 컨텍스트 식별자, 성능(예를 들어, 프로세싱 시간, 정확도), 모델 크기 등의 정보 중 적어도 하나가 포함될 수 있다. 추가로, 조합할 모델의 계층 정보로서 각 모델 또는 조정/전이 계층별로 가져올 모델의 식별자, 조합에 사용할 계층의 범위(모델 계층 범위), 또는 대상 모델 계층의 크기와 수신 정보(예를 들어, URL 등) 중 적어도 하나가 더 포함될 수 있다.

[0112] 서비스 애플리케이션은 단말로부터 요청 또는 수신된 전처리 정보로부터 판단된 컨텍스트에 대해 조합 모델 정보를 조회하고, 단말에서 보유하고 있는 모델의 재활용 가능 여부를 판단하고, 단말에 전송되어야 할 조합 계층 정보에서 모델 계층의 크기를 확인하고, 단말에서 조합될 때 제공할 수 있는 정확도, 프로세싱 시간 등을 확인할 수 있다.

[0113] 조회 및 가능 여부 판단 결과, 하나 이상의 조합 모델들이 가용한 경우, 서비스 우선 순위에 따라 하나 또는 복수 개의 조합 모델 정보를 선정하고 이를 단말에 전송할 수 있다. 예컨대 전송되어야 할 데이터의 크기가 가장

작은 것이 가장 높은 우선 순위를 가지는 실시예에서, 선정될 조합 모델 정보에는 단말에 전송되어야 할 모델 계층의 크기가 가장 작은 것이 포함될 수 있다.

[0114] 조합 모델과 관련된 조합 모델 구성 정보의 요소는 아래의 표 9와 같이 나타낼 수 있다:

표 9

[0115]

이름	의미
조합 계층 정보()	
- 모델 식별자	부분적으로 계층을 참고해 사용할 부분 모델의 식별자
- 모델 버전	모델의 버전
- 모델 계층 범위 (n,m)	계층의 범위: n번 계층부터 m번 계층까지, 또는 계층 식별자 n으로부터 계층 식별자 m까지임을 표시
- 모델 계층 식별자 목록	모델의 계층별 식별자들의 목록
- 모델 계층 버전 목록	모델의 계층별 버전 목록
- 모델 계층 크기	n번 계층부터 m번 계층까지의 파일 크기
- 모델 계층 URL	n번 계층부터 m번 계층까지에 해당하는 데이터를 수신할 수 있는 주소

[0116] 단말 애플리케이션은 서비스 애플리케이션으로부터 수신된 모델 정보에 부가하여 조합 모델 구성 정보가 수신되는 경우, 단말에 이미 수신되어 있던 모델의 정보를 기반으로 재활용이 가능한 모델의 유무, 수신해야 할 계층의 크기, 조합된 모델의 정확도 등의 우선 순위에 따라 일 조합 모델을 결정하고 모델 전체 또는 모델의 일부 계층만을 수신하여 모델을 구성할 것을 결정할 수 있다.

[0117] 단말 애플리케이션은 전체 또는 일부 계층의 수신을 위해 단말의 모델 수신부로 하여금 네트워크의 모델 전송부와 전송 세션을 체결하도록 지시할 수 있다.

[0118] 단말의 모델 수신부는 단말 애플리케이션에서 사용하기로 결정한 모델 조합 정보를 수신하고, 수신해야 할 모델 계층을 식별하고, 모델 계층 URL로부터 해당 모델 계층 범위를 수신하고, 수신되어 있던 모델과 결합하여 새로운 조합 모델을 생성할 수 있다.

[0119] 단말의 모델 수신부는 생성된 조합 모델을 추론 엔진에 전달하고, 추론 엔진은 새로운 추론을 개시할 수 있다.

[0120] 3GPP TR 26.927의 Common AI model information에는 모델의 전반적인 속성 정보가 포함된다. 단말에서 조합 모델을 생성하기 위한 조합 관련 정보는 아래의 표 10과 같이 나타낼 수 있다.

표 10

[0121]

Metadata category	Metadata type	Definition	Metadata type description (Examples)

Model information	Model identifier	An identifier for an AI model (or variants of it) specified for a certain AI media service. The identifier may be a name, a number, a combination thereof, a hash value. The identifier is defined during the configuration stage.	model_1, model_2
	Number of parameters	Total number of parameters in the neural network.	11 million
	Model size	The size of the AI model file in megabytes.	40MB
	Input size	The maximum size of the input data supported by the AI model in kilobytes.	256 KB
	Output size	The maximum size of the output data supported by the AI model in kilobytes.	256 KB
	Accuracy	The trained accuracy of the AI model as a percentage.	85%
	Target inference latency	The target inference latency specified for a given AI model in milliseconds. Such latency is measured between the input and output layers of the AI model at inference. This value is related to the service inference latency requirement of the service for which the AI model is provided, as well as the typical hardware capabilities of an entity performing the inference of the model.	20ms
	Format/ framework	The format or framework used to express the AI model, including its version number.	Pytorch 2.0 ONNX 1.15.0
	Processing capabilities	Estimated capabilities for processing the model including the computational power such as the computational cost (in FLOPS), the computational complexity (in MAC operations). It also includes the temporary memory to store model parameters.	NPU 10TFLOPS, MEM 10GB
Model composition	Referenced model identifier	An identifier of model which to be referenced.	Model identifier
	Model layer identifier	A list of identifiers of layers, or order number of layers of the referenced model.	[1:3], which means from layer 1 to 3 OR [layer_identifer1:layer_identifier3]
	Model layer size	The size of model layer(s) to be referenced.	10MB, 500KB
	Model layer URL	URL to download specific layer from network	URL
	Type of model composition	Method used for model performance improvement	fine_tuning OR transfer_learning

[0122] 표 10: Common AI model information

[0123] 표 10에서, Referenced model identifier는 참고 되어 조합의 구성이 될 대상 모델의 식별자를 나타낸다.

[0124] Model layer identifier는 모델 계층 범위 또는 모델 계층 식별자 목록일 수 있다. 참고될 모델의 계층을 n번째부터 m번째까지, 예컨대 [n,m]과 같이 범위로서 표시할 수도 있고, 각 계층의 각 식별자를 나열할 수도 있다.

[0125] Model layer size는 참조될 모델 계층 각각 또는 범위에 대한 전송 데이터 크기를 나타낸다.

[0126] Model layer URL은 참조될 모델 계층 각각 또는 범위에 대한 전송을 위한 데이터 주소를 나타낸다.

[0127] 도 2는 본 개시의 일 실시예에 따른 네트워크 엔티티(200)를 도시한 도면이다. 일 실시예에서, 네트워크 엔티티(200)는 전술한 도 1의 단말(UE) 또는 네트워크(서비스 제공자, 서버 등)에 대응될 수 있다.

[0128] 도 2를 참조하면, 네트워크 엔티티(200)는 송수신부(210), 프로세서(220), 및 메모리(230)로 구성될 수 있다. 전술한 네트워크 엔티티(200)의 통신 방법에 따라, 네트워크 엔티티(200)의 송수신부(210), 프로세서(220), 및 메모리(230)가 동작할 수 있다. 다만, 네트워크 엔티티(200)의 구성 요소가 전술한 예에 한정되는 것은 아니다. 예를 들어, 네트워크 엔티티(200)는 전술한 구성 요소들 보다 더 많은 구성 요소를 포함할 수도 있다. 일 실시

예에서, 송수신부(210), 프로세서(220), 및 메모리(230)는 하나의 칩(chip) 형태로 구현될 수도 있다. 또한, 프로세서(220)는 하나 이상의 프로세서를 포함할 수 있다.

[0129] 송수신부(210)는 네트워크 엔티티(200)의 수신부와 네트워크 엔티티(200)의 송신부를 통칭한 것으로서, 단말, 네트워크, 또는 기지국과 신호를 송수신할 수 있다. 단말, 네트워크, 또는 기지국과 송수신하는 신호는 제어 정보 및 데이터를 포함할 수 있다.

[0130] 또한, 송수신부(210)는 무선 채널을 통해 신호를 송수신하기 위한 기능들을 수행할 수 있다. 예를 들어, 송수신부(210)는 무선 채널을 통해 신호를 수신하여 프로세서(220)로 출력하고, 프로세서(220)로부터 출력된 신호를 무선 채널을 통해 전송할 수 있다.

[0131] 메모리(230)는 네트워크 엔티티(200)의 동작에 필요한 프로그램 및 데이터를 저장할 수 있다. 또한, 메모리(230)는 네트워크 엔티티(200)에서 획득되는 신호에 포함된 제어 정보 또는 데이터를 저장할 수 있다. 메모리(230)는 롬(ROM), 램(RAM), 하드디스크, CD-ROM 및 DVD 등과 같은 저장 매체 또는 저장 매체들의 조합으로 구성될 수 있다. 또한, 메모리(230)는 별도로 존재하지 않고 프로세서(220)에 포함되어 구성될 수도 있다. 메모리(230)는 휘발성 메모리, 비휘발성 메모리 또는 휘발성 메모리와 비휘발성 메모리의 조합으로 구성될 수 있다. 그리고, 메모리(230)는 프로세서(220)의 요청에 따라 저장된 데이터를 제공할 수 있다.

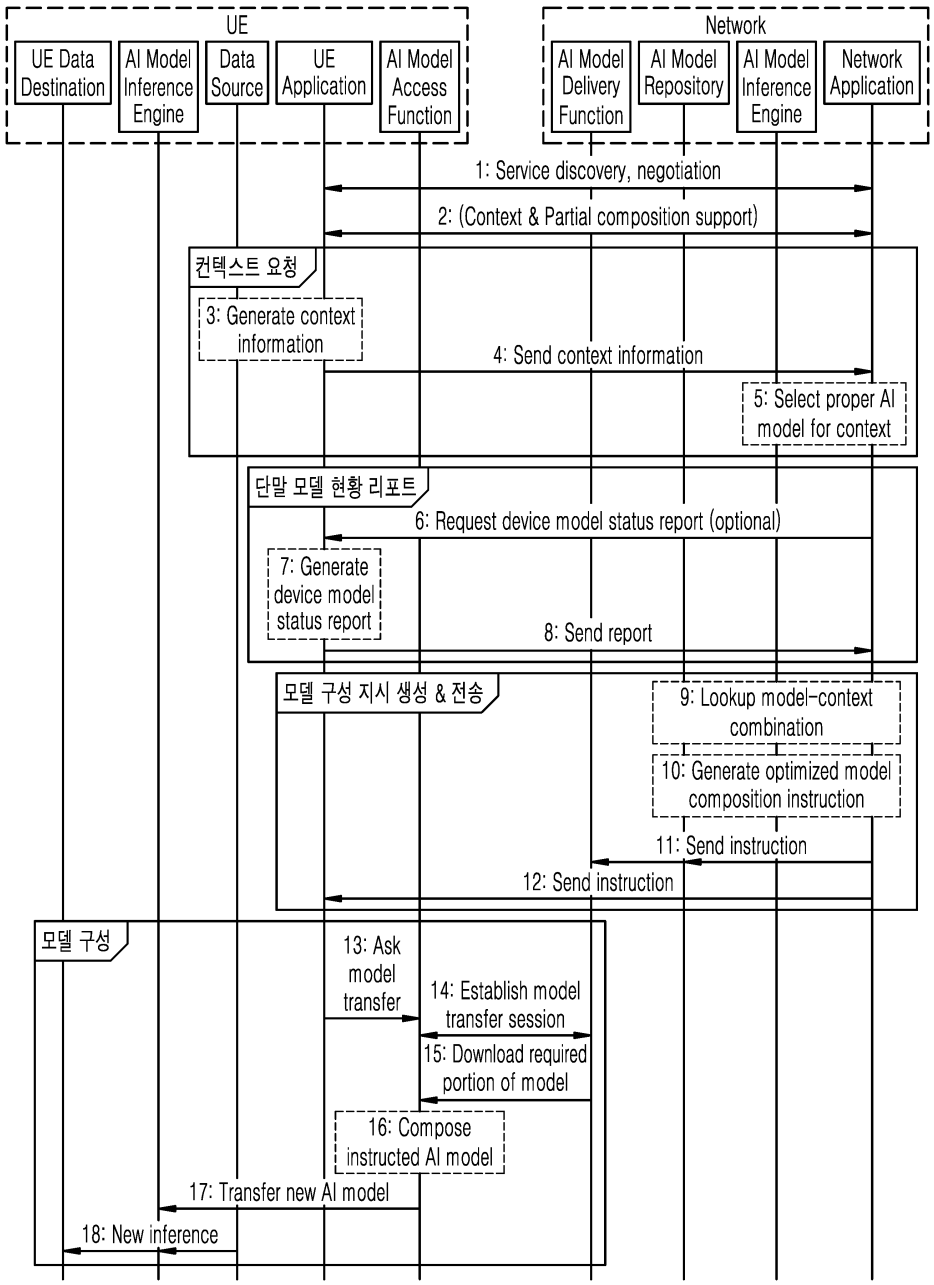
[0132] 프로세서(220)는 상술한 본 개시의 실시예에 따라 네트워크 엔티티(200)가 동작할 수 있도록 일련의 과정을 제어할 수 있다. 예를 들면, 프로세서(220)는 송수신부(210)를 통해 제어 신호와 데이터 신호를 수신하고, 수신한 제어 신호와 데이터 신호를 처리할 수 있다. 프로세서(220)는 처리한 제어 신호와 데이터 신호를 송수신부(210)를 통해 송신할 수 있다. 또한, 프로세서(220)는 메모리(230)에 데이터를 기록하거나 읽을 수 있다. 프로세서(220)는 통신 규격에서 요구하는 프로토콜 스택의 기능들을 수행할 수 있다. 이를 위해, 프로세서(220)는 적어도 하나의 프로세서 또는 마이크로(micro) 프로세서를 포함할 수 있다. 일 실시예에서, 송수신부(210)의 일부 또는 프로세서(220)는 CP(communication processor)로 지칭될 수 있다.

[0133] 전술한 본 개시의 설명은 예시를 위한 것이며, 본 개시가 속하는 기술분야의 통상의 지식을 가진 자는 본 개시의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[0134] 본 개시의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 개시의 범위에 포함되는 것으로 해석되어야 한다.

도면

도면1


UE
UE Data Destination
AI Model Inference Engine
Data Source
UE Application
AI Model Access Function
Network
AI Model Delivery Function
AI Model Repository
AI Model Inference Engine
Network Application
1: Service discovery, negotiation
2: (Context & Partial composition support)
컨텍스트 요청
3: Generate context information
4: Send context information
5: Select proper AI model for context
단말 모델 현황 리포트
6: Request device model status report (optional)
7: Generate device model status report
8: Send report
모델 구성 지시 생성 & 전송
9: Lookup model-context combination
10: Generate optimized model composition instruction
11: Send instruction
12: Send instruction
모델 구성
13: Ask model transfer
14: Establish model transfer session
15: Download required portion of model
16: Compose instructed AI model
17: Transfer new AI model
18: New inference

도면2