

AI/ML 서비스 지원 위한 AIoT 플랫폼의 머신러닝 데이터셋 관리 인터페이스

정승명*, 최충재

*한국전자기술연구원, 한국전자기술연구원

*sm.jeong@keti.re.kr, cjchoe1@keti.re.kr

AI/ML dataset management interface for AIoT platform to support AI services

SeungMyeong Jeong*, ChungJae Choe

*Korea Electronics Technology Institute. Korea Electronics Technology Institute

요약

AI와 IoT를 접목한 AIoT의 개념과 같이 IoT 데이터를 활용한 AI/ML 서비스가 확산되고 있다. 본 논문은 AI/ML 서비스 개발 시 IoT 센서에서 수집한 데이터로 모델 학습 및 모델 추론에 필요한 데이터셋을 생성하여 제공하는 인터페이스를 제안한다. 제안하는 인터페이스는 기존 IoT 미들웨어 플랫폼 표준이면서 AI 지원 기능을 추가 개발하고 있는 oneM2M 표준의 기존 인터페이스와 연계하여 확장하는 방안으로 설계되었다. 이를 통해 oneM2M 플랫폼으로 수집하는 다양한 IoT 데이터를 결합하여 학습용 데이터셋을 oneM2M 플랫폼이 생성해주고, 이를 조회하여 모델 생성이 가능하다. 이후 생성된 모델을 서비스에 배포하여 실행할 때 필요한 추론 입력 데이터 또한 동일한 IoT 데이터 속성을 포함한 데이터를 실시간으로 생성하여 AI 서비스 애플리케이션에 제공할 수 있다.

I. 서론

본 논문에서는 AI 학습 모델을 활용한 AI 서비스 개발에 필요한 데이터셋을 생성 관리하는 인터페이스를 제안한다. IoT 시스템으로 수집된 다량의 데이터가 확보되면 AI를 활용한 자동화된 서비스 개발이 용이하다[1]. 이미지 등 딥러닝이 아닌 AI 학습 모델 생성을 위해서는 JSON과 같은 텍스트 기반의 과거 시계열 데이터가 필요하며 특히 AIoT 융복합 서비스에서는 복수의 IoT 센싱 이력 데이터를 활용한 데이터셋 준비가 필요하다.

제안하는 방법은 IoT 국제표준인 oneM2M 기술을 대상으로 하고 있으며 기존 oneM2M 플랫폼은 센서 별로 다른 리소스에 데이터를 저장한다. 따라서 AI 학습 모델 개발자는 각각 센서의 이력 데이터를 조회하여 데이터 생성 시간과 같은 시간 정보를 기준으로 통합된 데이터셋을 생성한다.

데이터를 표준 인터페이스로 개별 조회하여 데이터셋을 생성하는 과정은 상대적으로 시간도 오래 소요될뿐만 아니라, IoT 장치 및 데이터를 관리하는 oneM2M 표준 입장에서 데이터셋 관리를 플랫폼 공통 기능으로 제공할 경우 애플리케이션 입장에서 그 활용성이 높다고 볼 수 있다.

학습용 데이터셋을 oneM2M 플랫폼에서 제공해주는 것뿐만 아니라 개발자가 AI 모델 생성 이후 이를 실행할 때 필요한 추론 입력 데이터 생성 또한 플랫폼에서 제공할 수 있다. 이 때는 기존 이력 데이터가 아닌 신규 생성되는 데이터들을 결합하여 추론용 데이터를 생성하여 저장하고 애플리케이션에 실시간으로 제공한다.

대상 기술로서 oneM2M은 분산 플랫폼을 지원하는 아키텍처로 구성되어 있어 수집된 프라이버시가 포함된 데이터를 외부로 제공하지 않고 Edge에서 모델 추론에 필요한 데이터를 사용할 수 있는 장점이 있다. 또한 Edge 디바이스를 활용한 AI 서비스는 필드에서 수집한 데이터를 매번 서버에 전송하여 서버에서 학습과 추론을 중앙 집중식으로 수행하지 않고 분산하여 처리하는 장점 또한 가지고 있다[2].

이전 연구에서는 학습 데이터셋은 조회 요청의 응답으로 회신하고 추론 데이터를 리소스로 생성하여 AI 애플리케이션에 제공하였다. 그리고 빈

(null) 값 처리 방법 등 데이터셋 생성에 필요한 정책 정보를 제안하였다. 본 논문은 해당 정책 정보를 활용하되 학습 데이터셋과 추론 데이터 생성 시 같은 데이터 속성, 전처리 등을 구형해야하는 점을 고려하여 하나의 정책 리소스로 두 데이터셋을 관리하는 차이점을 가진다.

II. 본론

그림1은 본 논문에서 제안하는 oneM2M 플랫폼을 활용한 AI 데이터셋 생성 관리 기능을 활용한 AIoT 시스템 동작 과정을 개략적으로 나타낸다. oneM2M 표준 인터페이스로 수집된 데이터에 대해 AI 개발자가 데이터셋 생성을 요청하고 이를 통해 생성된 데이터셋을 조회하여 AI 모델을 생성하고 이를 서비스 애플리케이션에 배포한다. 이후 개발자의 구독 생성에 의해 새로운 센서 데이터가 플랫폼에 저장될 때마다 플랫폼이 추론 입력 데이터를 통지 메시지로 애플리케이션에 제공하여 모델 추론에 의한 AI 예측 서비스를 수행한다.

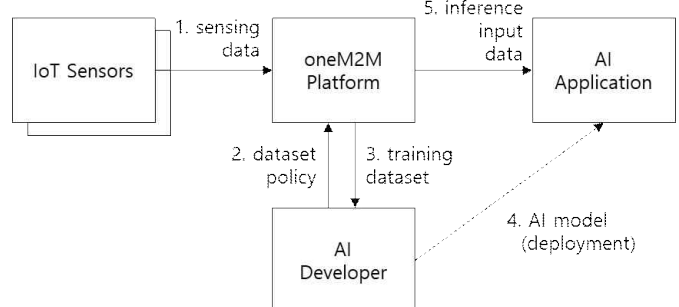


그림1 ML 데이터셋 생성 관리 기능 활용 시스템 예시

제안하는 머신러닝 데이터셋 관리 인터페이스는 REST API를 채용한 oneM2M 표준의 특성을 반영하여 신규 리소스 및 관련 프로시저로 정의한다. oneM2M 표준은 애플리케이션은 플랫폼에 기능 수행 요청 시 요청하는 기능에 해당하는 리소스를 생성한다. 예를 들어 다중 리소스에 대한 일괄 접

근을 위해서 <group> 리소스를 생성하고 이후 생성된 리소스를 활용해 일괄 조회 등 oneM2M 플랫폼이 제공하는 기능을 이용한다.

마찬가지로 기존에 oneM2M 플랫폼에 저장된 또는 지속적으로 신규 데이터를 저장하고 있는 데이터 리소스를 병합하는 요청 또한 그림2의 리소스 구조 예시와 같이 <datasetPolicy> 리소스를 플랫폼에 생성하는 것으로 요청할 수 있다. 이 리소스는 이름과 같이 데이터셋을 생성하고 관리하기 위한 정책 및 설정 정보를 가지고 있다. 예시에서 datasetPolicy 타입인 policy1 리소스는 원본 데이터 리소스 식별자를 가지며 base/parking 및 base/weather 리소스를 가리킨다.

애플리케이션 요청에 의해 policy1 리소스를 생성하면 플랫폼은 학습 데이터는 <dataset> 리소스를 생성하고 자녀 리소스로 실제 데이터셋을 <datasetFragment> 리소스로 저장한다. 이후 <datasetPolicy> 리소스에 학습 데이터셋 리소스 식별자를 저장하고, 정책 리소스 식별자를 데이터셋 리소스에 policyID 속성으로 저장한다.

<dataset> 리소스가 데이터셋의 메타 데이터를 저장하는 반면 실제 데이터셋은 정책 설정에 따라 여러 리소스로 분할되어 자녀 리소스로 저장된다. 그림2에 도시된 예시와 같이 학습 데이터셋은 원본 리소스의 기존(이력) 데이터를 병합하여 생성된다. 반면 추론용 데이터는 같은 원본 리소스에 같은 정책 정보를 적용하여 원본 리소스에 신규 생성되는 데이터를 병합하여 추론 데이터셋 리소스로 저장한다.

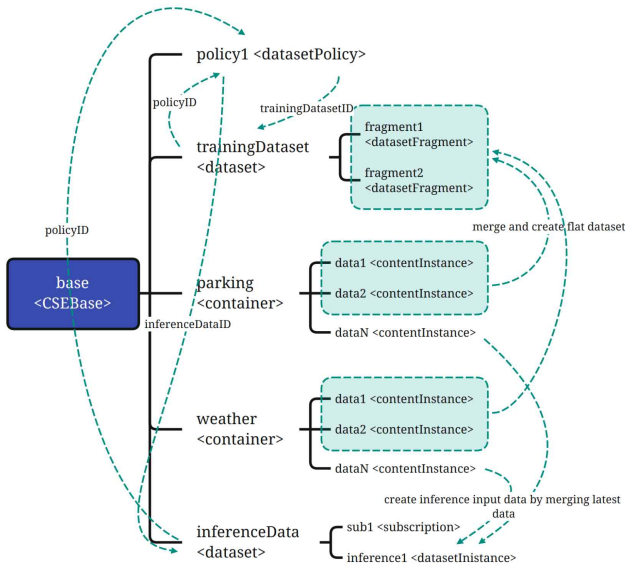


그림2 데이터셋 생성 관리를 위한 oneM2M 리소스 구조 설계

표1은 상기 리소스 및 인터페이스를 통해 생성되는 학습 데이터셋 예시를 나타낸다. 주차장 데이터에서 가용주차면 수를 나타내는 availableNo 속성과 날씨 데이터에서 온도 및 기상현황 속성을 데이터셋으로 생성하였고 이 때 각 <contentInstance> 리소스의 creationTime(ct) 속성 값을 기준으로 데이터를 병합하였다.

표1 다중 센서 데이터 병합한 학습 데이터셋 예시

| ct | availableNo | temperature | weatherType |
|-----------------|-------------|-------------|-------------|
| 20220901T000000 | | 18.4 | "cloudy" |
| 20220901T000303 | 17 | | |
| 20220901T000505 | 16 | | |
| 20220901T010000 | | 15.6 | "rainy" |
| 20220901T010012 | 15 | | |

추론 데이터는 원본 리소스에 신규 데이터가 생성될 때마다 생성된다. 따라서 그림2의 예시와 같이 추론 데이터셋에 구독(<subscription> 리소스)을 신규 자녀 리소스 생성 이벤트에 대한 통지로 정책을 설정하면 AI 모델을 구동하는 애플리케이션에 실시간 추론용 데이터를 통지 메시지를 통해 제공할 수 있다[4].

주차장의 availableNo 속성은 주차 이벤트 발생에 따라 비주기적으로 생성되는 데이터이며 날씨 관측 데이터는 주기적으로 생성된다. 추론 입력 데이터는 학습 데이터셋의 모든 속성(feature)값을 포함한다. 따라서 추론 데이터는 이벤트 데이터의 경우 마지막 이벤트 데이터를 추론 데이터에 포함시키며, 주기 데이터의 경우 사용자가 원하는 정책에 따라 마지막 관측값 또는 평균 등의 값으로 신규 추론 입력 데이터에 포함할 수 있다. 표2는 정시에 기상 관측 데이터가 플랫폼에 신규 리소스로 저장될 때 플랫폼이 생성한 추론 입력 데이터 예시를 나타낸다. 해당 시각에 주차장 이벤트는 발생하지 않았지만 마지막으로 보고된 여유 주차면 수를 추론 입력 데이터에 포함하여 <datasetFragment> 리소스를 생성한다. 사전에 구독이 생성되어 있으면 ML 애플리케이션에 아래 데이터를 포함한 리소스가 통지 메시지로 전송된다.

표2 다중 센서 데이터 병합한 추론 입력 데이터 예시

| ct | availableNo | temperature | weatherType |
|-----------------|-------------|-------------|-------------|
| 20220902T100000 | 4 | 23.4 | "sunny" |

III. 결론

본 논문에서는 IoT 표준 기술인 oneM2M에 AI/ML 데이터셋 생성 관리 인터페이스를 설계하여 oneM2M 플랫폼이 생성하는 학습 데이터셋을 활용해 AI 모델을 생성하고 이 모델을 탑재한 애플리케이션에 실시간 추론 데이터를 제공하여 AI 기반 서비스를 구축할 수 있는 방안을 제시하였다.

추후 제안 기술을 oneM2M의 분산 플랫폼 구조와 접목하여 서버에서 기본 모델을 디바이스로 배포하고 디바이스 별로 모델을 고도화하여 서버에서 갱신된 모델 정보를 병합하는 연합학습(federated learning)에도 적용할 수 있다. 개인 정보가 서비스 제공자의 서버로 유출되지 않고 AI 기반 서비스를 사용하기 위한 시스템 구조를 oneM2M에서 지원하고 있으므로 본 논문에서 제안하는 기술을 바탕으로 후속 연구 수행할 예정이다.

ACKNOWLEDGMENT

이 연구는 2022년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임(20016243)

참 고 문 헌

- [1] Zhala J. H., et al, "Machine Learning Powered IoT for Smart Applications," International Journal of Science and Business, 5(3), 92-100.
- [2] Adi, Erwin, et al, "Machine learning and data analytics for the IoT." Neural Computing and Applications 32.20 (2020): 16205-16233.
- [3] Jeong, S., et al, "Machine Learning dataset creation common service for real-time data platform," International conference on Internet 2021,
- [4] oneM2M, "TS-0001, Functional Architecture TS v4.0.0".