효율적 기계학습 모델 구축지원을 위한 JupyterLab 기반의 AI 워크벤치 개발

김성찬[1,2], 장래영[1], 최명석[1]

[1]한국과학기술정보연구원 기계학습데이터연구단, [2]과학기술연합대학원대학교 응용AI

sckim@kisti.re.kr, raezero@kisti.re.kr, mschoi@kisti.re.kr

Development of AI Workbench based on JupyterLab for Efficient Machine Learning Model Building

Seongchan Kim[1,2], Rae-young Jang[1], Myung-seok Choi[1]

[1]Dept. of Machine Learning Data Research, Korea Institute of Science and Technology Information

[2]Applied AI, UST-KISTI

요 약

본 논문은 효율적인 기계학습 모델 구축을 위한 JupyterLab 기반의 AI 워크벤치 개발에 관하여 기술한다. AI 워크벤치에서는 웹에서 JupyterLab 기반의 개발환경을 사용자별로 제공하여 데이터 관리 및 모델구축이 쉽도록 사용자 편의성을 극대화한다. AIDA(AI Data Archive)와의 연동을 통해 공개된 기계학습데이터와 사전학습 모델 등을 제공하여 필요 리소스의 접근이 쉽도록 한다. AI 워크벤치는 Kubernetes 기반으로 사용자별 개발환경을 컨테이너 기반으로 구성하여 운영하도록 구현하였으며 기계학습 모델의 지원을 위해 GPU 또한 활용할 수 있도록 하였다. 이로써 AI 모델 개발자들의 작업 효율성을 높이도록 하였다.

Ⅰ. 서 론

최근 데이터의 공개가 활발해 지면서 기계학습 모델 개발 또한 활발하게 이루어지고 있으며 이를 개발할수 있는 환경에 대한 수요도 늘어나고 있다. 또한 최근에는 사전학습 모델의 공개도 활발하게 이루어져 AI 모델 개발의 주요 방법론으로 부상하고 있다. 이러한 개발 방법은 이미지나 텍스트의 대규모의 데이터셋을 가지고 Self-supervised Learning 기법으로 사전에 뉴럴 네트워크를 학습하고, 추후에 사용자의 태스크에 맞게 새로운 데이터셋으로 전이학습(Transfer Learning)을 수행하며 사용자 태스크에 성능을 극대화 하는 것을 말한다.

한편 과학기술 데이터를 활용하여 AI 모델을 만들고 신물질 개발, 난치병 치료 등과 같은 과학기술 발전을 꾀하는 데이터 기반 과학 연구가 활발하게 진행되고 있으며 한국과학기술정보연구원에서는 이러한 흐름에 맞추어 기계학습 데이터를 공개하는 KISTI 기계학습 데이터 공유·활용 서비스 AIDA(AI Data Archive)[1]를 개발하여 서비스를 제공하고 있다.

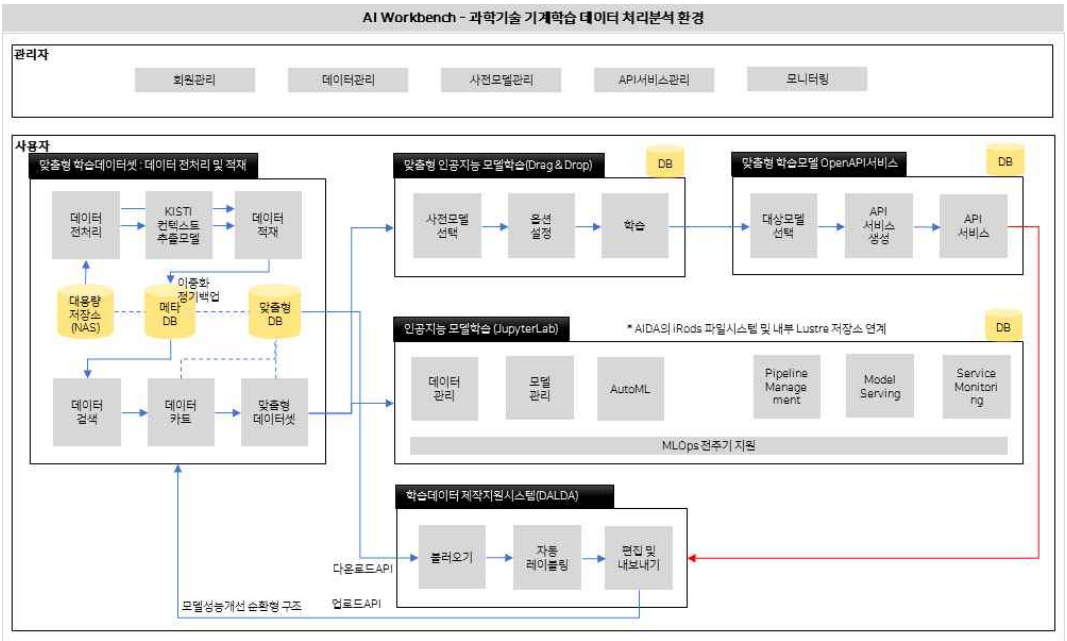

그림 1 AI 워크벤치 구조 및 MLOps 전주기 지원 JupyterLab 개발환경

AIDA는 과학기술 논문 보고서에서 추출한 텍스트, 표/그림 뿐 아니라 기관식별 데이터, 논문 QA데이터, 논문 의미태깅 데이터와 대전시 도로영상 등과 같은 다양한 형태의 데이터를 제공하고 있다. 또한 대규모의 과학기술 텍스트를 이용하여 사전 학습한 KorSciBERT, KorSciELECTRA와 같은 과학기술 사전학습 언어모델도 제공하고 있다. 이밖에 기상데이터, 네트워크 침입데이터와 같은 다양한 도메인의 데이터도 제공하고 있다.

Ⅱ. JupyterLab 기반의 AI 워크벤치 구축

데이터・사전학습 모델 등의 서비스 뿐 아니라 한국과학기술정보연구원은 과학기술 기계학습데이터 처리・분석을 지원하는 플랫폼인 AI 워크벤치 또한 개발하고 있다. 공개된 데이터를 바로 분석하기 쉽게 지원하도록 하는 프레임워크나 서비스들이 최근 많이 등장하고 있으며, 클라우드 기반의 연구데이터 분석을 지원하는 CANVAS [2], 기상기후 데이터 분석을 지원하는 기상기후 빅데이터 분석플랫폼 날씨마루 [3]와 같은 서비스들이 대표적인 예라 할 수 있다. AI 워크벤치의 주요 목표 및 다른 서비스들과의 차이점은 맞춤형 데이터셋 기반의 학습을 지원한다는 것과 MLOps 전주기 지원의 JupyterLab 기반의 통합 개발환경을 지원하는 것이며, 다음과 같은 주요 컴포넌트 개발을 목표로 하고 있다.

- 대용량 학습데이터 및 사전학습모델 통합 관리
- 직관적 UI 기반 맞춤형 사전학습 모델 기반 모델 훈련
- JupyterLab 기반 AI 모델 생성 및 활용을 위한 웹기반 시스템

위 기술된 서비스를 제공하기 위하여 웹서비스를 구현하여 제공하고 다중사용자에게 안정적인 처리 분석환경을 제공하고자 컨테이너 기반의 자원관리 오픈소스인 Kubernetes를 설치하여 연산자원을 관리한다. 또한, JupyterLab 기반의 AI 워크벤치는 다음의 주요 컴포넌트를 구축한다.

- 웹 기반 모델 훈련 워크플로우 통합 UI 개발
- 사전학습 모델 기반의 전이학습 지원 기능 개발
- 사용자 및 데이터 관리 기능 개발
- MLOps 전주기 지원 기능 개발

웹상에서 JupyberLab UI를 기반으로 모델을 작성하고 훈련시킬 수 있는 환경을 제공하고 사용자 데이터를 저장할 수 있는 저장공간(e.g., 100GB)을 제공하여 AIDA에서 제공하는 사전학습 모델과 데이터는 IRODS 커넥터를 이용하여 IRODS상에 저장된 모델과 데이터를 다운로드 없이 활용할 수 있도록 설계 하였다. 사용자는 사전학습 모델을 지정하고 사용자의 데이터를 이용하여 전이학습을 수행할 수 있다. 이때 생성된 모델과 데이터는 AIDA와 연동되어 손쉽게 공개할 수 있도록 구축되었으며 데이터와 모델 관련하여 활용 권한 제어를 할 수 있도록 하였다.

Ⅲ. 결론

본 논문에서는 과학기술 기계학습데이터 처리・분석을 지원하는 플랫폼인 AI 워크벤치의 웹기반의 JupyterLab 개발에 관하여 기술하였다. 추후 연구로 사용자의 데이터수집, 모델개발, 모델운영 및 API서비스 구축에 이르는 기계학습 모델 서비의의 전주기 지원을 위하여 Kubeflow[4]와 같은 프레임워크를 도입하여 서비스화 하는 것을 고려하고 있다. 또한 기계학습 모델 구축시 적절한 자질을 선택하도록 하는 플러그인과 코드의 결함을 자동으로 찾아내어 사용자를 지원하는 기능의 탑재를 고려하고 있다.

ACKNOWLEDGMENT

이 연구는 2022년도 한국과학기술정보연구원의 기본사업 “Data/AI 기반 문제해결체계 구축”(K-22-L04-C05-S01)의 지원으로 수행되었습니다.

참 고 문 헌

[1] KISTI 기계학습 데이터 공유·활용 서비스 (https://aida.kisti.re.kr)

[2] 김성찬, 송사광, "CANVAS: 클라우드 기반 연구데이터 분석 환경 및 시스템", 2021, 한국컴퓨터정보학회논문지 vol.26, no.10, 통권 211호 pp. 117-124

[3] 기상기후 빅데이터 분석 플랫폼, (https://bd.kma.go.kr/)

[4] Kubeflow, (https://www.kubeflow.org/).