국가연구데이터커먼즈를 활용한 딥러닝 학습 모델 접근성 향상을 위한 재현 방안

이상백 [1], 김다솔 [1], 송사광 [1], 조민희 [1], 이미경 [1], 임형준 [1]
[1] 한국과학기술정보연구원 연구데이터공유센터

sb_lee@kisti.re.kr, dskim@kisti.re.kr, esmallj@kisti.re.kr, mini@kisti.re.kr, jerryis@kisti.re.kr, hjlim@kisti.re.kr

Reproducibility Approach for Enhancing Accessibility of Deep Learning Models Using the Korea Research Data Commons

Sang-baek Lee[1], Dasol Kim[1], Sa-kwang Song[1], Minhee Cho[1], Mikyung Lee[1], Hyung-Jun Yim[1]
[1]Research Data Sharing Center, Korea Institute of Science and Technology Information(KISTI)

요 약

딥러닝에 대한 관심이 증가함에 따라 다양한 분야의 연구자 사이에 딥러닝 모델의 적용 및 재현이 중요한 작업으로 자리잡았다. 하지만 모델을 재현하고 활용하는데 있어 다양한 환경과 자원의 한계가 발생하여 문제가 되고 있다. 이러한 문제를 해결하기 위해 본 논문에서는 국가연구데이터커먼즈체계인 KRDC 프레임워크를 활용하여 딥러닝 학습 모델의 재현 방안을 제안하였다. 이를 통해 딥러닝 연구에 익숙하지 않은 사용자도 학습 모델의 적용 및 활용을 용이하게 할 수 있음을 확인하였다. KRDC 프레임워크는 사용자가 원하는 데이터와 태스크를 정의하고, 워크플로우로 구성, 학습 모델의 재현 및 활용을 지원한다.

1. 서론

최근 딥러닝에 대한 관심이 폭발적으로 증가함에 따라 기존의 기계학습과 관련된 연구자 뿐만 아니라 다른 분야의 연구자나 일반시민연구자에 이르기까지 다양한 사람들이 딥러닝 모델을 도입하여 새로운 시도를 하고 있다. 꼭 기계학습 분야 연구자가 아니더라도 컴퓨터공학에 익숙한 연구자의 경우에는 깃헙과 같은 오픈소스 저장소에 공유된 소스코드와 마크다운(.md)파일의 설명을 통해 학습 모델을 재현해보고, 본인의 데이터를 통해 공유된 학습 모델을 새롭게 학습시켜 활용하기도 한다. 대부분의 학습 모델에 관한 연구의 경우에 각각의 연구자들이 본인의 개발 환경에 맞춰 학습을 진행하게 된다. 개발 환경, GPU 종류, 하이퍼파라미터 설정 등 다양한 요소에 따라 학습 모델의 성능이 달리지기 때문에 연구자들은 최적의 학습 모델 성능을 달성하기 위해 다양한 실험을 하여야 한다. 이와 같은 반복적인 과정을 거쳐 최적의 학습 모델 성능을 달성하게 되면 이를 논문을 통해 발표하고 다른 연구자들과 공유하게 된다. 이 과정에서 발표된 학습 모델의 성능에 대한 검증을 위하여 학습 모델의 재현을 위한 학습 환경과 소스코드 등을 공유하는 것이 일반적이다. 연구자가 자신의 학습 모델에 대한 성능 검증을 위하여 재현할 수 있는 모든 정보를 공유한다고 해서 다른 연구자가 항상 똑같은 결과를 재현할 수 있는 것은 아니다. 앞서 설명한 바와 같이 개개인이 가지고 있는 개발 환경, GPU 환경 등

이 다르기도 하고, 기존 연구자에 비해 부족한 컴퓨팅리소스를 가지고 있다면 학습 결과가 달라질 수 있기 때문이다. 이와 같이 연구자의 학습 환경이 부족한 경우를 보완하기 위한 방법에는 가상 환경과 같은 클라우드 컴퓨팅을 이용하기도 한다. 구글 Colab, 아마존 AWS 등과 같은 클라우드 서비스를 이용하거나 다양한 MLOps 서비스를 이용하는 방법이 있다. 그러나 딥러닝 연구에 익숙하지 않거나 컴퓨터공학 분야의 연구자가 아닌 경우에 이와 같은 재현 방식들은 연구자들에게 진입 장벽으로 작용할 수 밖에 없다. 이에 본 논문에서는 국가연구데이터커먼즈체계인 KRDC 프레임워크 기반의 딥러닝 학습 모델 재현 방안에 대해 제안하고자 한다. 본 논문에서는 KRDC 를 이용하여 기존의 SOTA 성능을 달성하고 있는 학습 모델의 재현이 가능함을 보이고, 딥러닝 연구에 익숙하지 않은 사용자도 학습 모델의 적용 및 활용이 가능함을 보이고자 한다.

2. 배경 및 관련 연구

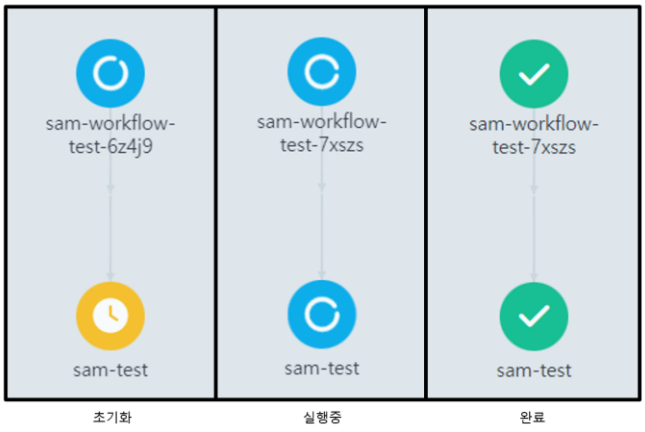
딥러닝 분야의 성장에 맞춰 다양한 분야에서 딥러닝 연구를 하기 위한 시도를 하고 있다. 이에 따라 기존의 컴퓨팅 자원을 활용하는 경우도 있지만, 대부분의 경우 딥러닝 연구를 수행할 수 있는 정도의 컴퓨팅 자원이 없기 때문에 새롭게 인프라를 구축해야 하는 어려움이 있다. 이에 최근에는 클라우드 컴퓨팅이나 MLOps 와 같은 연구 환경을 제공해주는 서비스의 필요성이 증가하고 있다. 일시적인 학습 모델의 실험이나 활용을 위해 연구 환경을 자체적으로 구축하는 것은 매우 큰 비용이 소요되는 작업이기 때문이다. 그러나 클라우드 컴퓨팅이나 MLOps 와 같이 인프라 자원을 제공받을 수 있는 환경을 갖췄다고 하더라도, 오픈소스를 수정하거나 활용할 수 없는 경우에는 공개된 학습 모델에 대한 활용이 어려운 경우가 발생한다.

국가연구데이터커먼즈(KRDC; Korea Research Data Commons)는 국가 차원에서 데이터 기반 연구개발 환경을 제공하기 위한 것으로 연구자가 KRDC 를 활용하여 다양한 연구데이터를 분석, 활용할 수 있다. 이와 같은 공통 체계를 딥러닝 연구와 같은 다소 진입장벽이 높은 분야에 활용한다면 기존의 딥러닝 연구자뿐만 아니라, 배경지식이 적은 다른 도메인의 연구자들도 손쉽게 학습 모델에 대한 적용, 활용을 할 수 있게 된다. 또한, KRDC 는 연구데이터 분석, 활용 기능 뿐만 아니라 소스코드 버전 관리, 소스 제어, 저장소, 협업, CI(Continuous Integration)/CD(Continuous Deployment) 등의 다양한 기능을 제공하고 있기 때문에 소프트웨어 개발 및 소스코드 구현에 대한 사전지식이 없는 사용자도 원하는 태스크에 맞춰 워크플로우를 구성하고 학습데이터를 활용할 수 있도록 환경을 제공한다.[1]

3. KRDC 프레임워크 기반 모델 재현

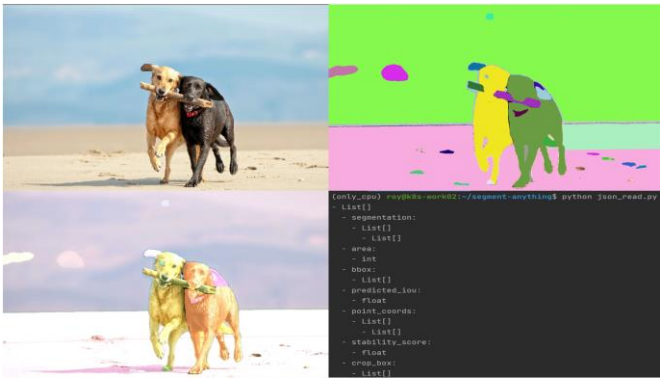
딥러닝 분야에 대한 연구를 희망하는 연구자들을 지원하기 위하여 국가연구데이터커먼즈 체계에서는 학습 모델을 재현하고 활용할 수 있는 환경을 제공하고 있다. 특정 도메인의 연구자가 학습 모델을 공유하기 위하여 소스코드를 국가연구데이터커먼즈 체계에 탑재를 하게 되면, 해당 학습 모델은 하나의 리소스로 공유, 활용이 가능한 형태로 저장이 된다. 이를 다양한 사용자가 본인이 원하는 형태의 데이터를 활용하여 학습 모델을 새롭게 학습시키거나, 기존의 학습이 완료된 가중치를 활용하여 추론 결과값을 얻을 수 있다. [그림 1]과 같이 공유된 학습 모델을 선택하여 태스크를 정의하고, 데이터를 입력하게 되면 워크플로우가 구성이 된다. 구성이 완료된 워크플로우는 실행이 되면 순차적으로 초기화를 진행하고, 실행이 되는데 이때 입력으로 주어진 데이터를 이용하여 학습을 하거나 추론을 하게 된다. 사용자가 원하는 태스크가 완료가 되면 워크플로우는 완료가 되고, 최초 정의된 형태의 출력을 얻을 수 있다. 또한, 연구데이터커먼즈 체계에 탑재한 리소스는 제공자의 승인 여부에 따라 다른 사용자로 하여, 수정, 활용을 할 수

있도록 권한을 부여할 수 있다. 딥러닝 학습 모델에 대한 사전 지식이 없는 사용자뿐만 아니라, 소스코드를 수정하여 기존의 학습 모델에서 향상된 학습 모델을 생성하고 싶은 사용자에게도 개발 환경 및 인프라를 제공할 수 있다.



(그림 1) 연구데이터커먼즈 워크플로우 구동 형태

본 논문에서는 페이스북의 Meta AI Research 에서 공개한 객체 분할(Object Segmentation) 연구 분야에서 SOTA(State-of-the-Art) 성능을 달성한 Segment-Anything[2] 학습 모델을 이용하여 국가연구데이터커먼즈를 통해 학습 모델 재현 및 활용이 가능함을 보였다. [그림 2]에서 확인할 수 있듯이 원본 이미지를 입력으로 넣었을 때, 사용자의 설정에 따라서 hard segmentation, blended segmentation 중에 원하는 형태의 이미지 출력을 선택하여 획득할 수 있다. 또한, 분할 결과를 다른 연구와 연계하여 사용하는 경우나 출력 값을 이용하여 새로운 정보를 만들어내고자 할 경우에는 출력 형태를 이미지가 아닌 JSON 으로 설정하여 입력 이미지에서 분할 정보가 어느 지점에 해당하는지를 JSON 형태로 얻을 수 있는 것도 확인하였다.



(그림 2) (top left) 원본, (top right) hard segment, (bottom left) blended segment, (bottom right) JSON 출력

4. 결론

본 논문에서는 딥러닝 모델의 활용과 재현이 중요해짐에 따라 국가연구데이터커먼즈를 통해 연구자와 개발자, 그리고 일반시민과학자 등을 포함하는 사용자에게 새로운 연구 환경과 자원을 제공함으로써 딥러닝 학습 모델의 쉬운 접근성과 활용성을 제공할 수 있음을 확인하였다. 특히, 복잡한 딥러닝 환경 및 자원 문제를 해결하기 위한 이러한 접근은 딥러닝 연구에 익숙하지 않은 사용자들이 모델을 쉽게 재현하고 활용할 수 있는 새로운 기회를 제공할 수 있음을 확인하였다. 본 논문을 통해 제안한 KRDC 기반의 딥러닝 학습 모델 재현 방안을 활용하여 딥러닝 연구 및 활용의 진입 장벽을 낮추는데 중요한 발판이 될 것으로 기대된다. 또한, 향후 연구로 다양한 연구분야에서 발생하는 연구데이터 및 연구소프트웨어 등을 공유, 활용할 수 있는 국가연구데이터커먼즈 체계를 보다 견고하게 확립함으로써 모든 연구자들이 다양한 연구분야에서 딥러닝을 활용할 수 있도록 접근성을 높이고, 협업의 효율성을 극대화하여 연구의 질과 속도를 향상시킬 수 있는 환경을 구축하는 것을 목표로 하고자 한다.

ACKNOWLEDGEMENT

이 논문은 2023 년도 한국과학기술정보연구원(KISTI)의 기본사업으로 수행된 연구입니다. (과제번호: (KISTI)K-23-L01-C03-S01, (NTIS)1711198423)

참고문헌

[1] 임형준, 이미경, 송사광, 서동민, 조민희. 데이터 기반 연구개발을 위한 국가연구데이터커먼즈 설계 및 적용 방안. 한국지능시스템학회 논문지, 32(5), 392-400, 2022

[2] Kirillov, Alexander, et al. "Segment anything." arXiv preprint arXiv:2304.02643 (2023).