신경망 자동생성 지원 MLOps 기술 동향

MLOps Technology Trend Supporting Automatic Generation of Neural Network

김선태 (S.T. Kim, stkim10@etri.re.kr) AI컴퓨팅시스템SW연구실 책임연구원
조창식 (C.S. Cho, chcho@etri.re.kr) AI컴퓨팅시스템SW연구실 책임연구원/실장

ABSTRACT

As more devices are used across various industries and their performance improves, artificial intelligence applications are being increasingly adopted. Hence, the rapid development of neural networks suitable for diverse devices can determine the competitiveness of companies. Machine learning operations (MLOps), which constitute a framework that supports neural network generation and its immediate application to devices, have become necessary for the development of artificial intelligence. Currently, most MLOps are provided by major companies such as Google, Amazon, and Microsoft, which provide cloud services supported by large-scale computing power. In addition, various services are provided by the open-source project Kubeflow. We examine basic concepts and technology trends in MLOps and unveil additional functions required in industry.

KEYWORDS Continual Deployment, Continual Integration, Generation of Neural Network, MLOps, Target-adaptive

Ⅰ. 서론

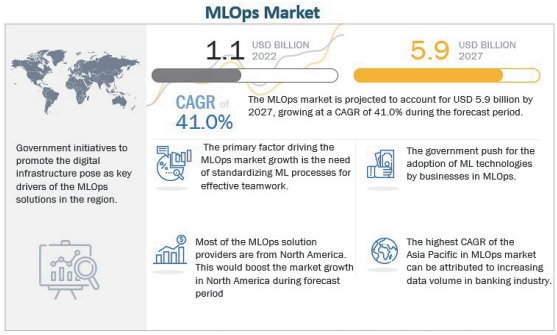
산업 전반에 적용되는 디바이스는 수많은 종류가 존재하며, 각각의 디바이스에 맞는 신경망을 별도로 생성해서 탑재하여야 하는 문제점이 있다. 이런 환경에서 업무의 효율성을 지원하는 인공지능 알고리즘의 산업적용이 제한되었다. 이러한 문제를 해결하기 위해서 AutoML(신경망 생성) 기술이 대두되었지만 주어진 몇 개의 디바이스 등급(Class)에 제한되어 적용하고 있다. 뿐만 아니라 개발된 신경망을 디바이스에 탑재할 경우, 디바이스에 맞는 최적 신경망 솔루션을 필요로 하였다. 그림 1의 머신러닝 워크플로우를 기본 요소로 하여, 지속적으로 유입되는 데이터에 대응하는 학습 및 업데이트, 그리고 개발 버전을 관리할 수 있는 모듈들이 추가되고 있다. 이런 모듈들의 총합을 MLOps(Machine Learning Operations)라고 한다.

본고에서는 산업환경에 적용되는 인공지능 응용을 위한 개발 및 관리를 지원하는 솔루션에 대해서 언급하고, 현재의 글로벌 기업들이 제공하는 관련

* DOI: https://doi.org/10.22648/ETRI.2024.J.390502
* 본고는 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임[NO. 2021-0-00766, 신경망 응용 자동생성 및 실행환경 최적화 배포를 지원하는 통합개발 프레임워크 기술개발].

출처 게티이미지뱅크, 무단 전재 및 재배포 금지

그림 1 머신러닝(ML)의 일반적 워크플로우



출처 Reprinted with permission from [1]. 도표의 저작권은 Market and Market에 있으며, Market and Market의 동의하에 사용되었습니다. 추후 이용 시 Market and Market에 문의하시기 바랍니다.

그림 2 MLOps 시장 규모 예측

솔루션에 대해서 알아보고자 한다. 뿐만 아니라 데이터를 보유하고 있는 산업환경에서 인공지능 전문가의 부재에도 자신의 환경에 맞는 신경망을 개발하고 관리할 수 있는 타겟 적응형 MLOps에 대해서 설명한다.

II. 주요 MLOps 동향

MLOps가 대두되면서 그림 2와 같이 시장 크기[1]가 2022년 11억 달러에서 2027년 59억 달러로 CAGR 41.0%의 급격한 성장이 예측되고 있다. 이는 이 시장 생태계의 다양한 핵심 기업들이 경쟁적으로 핵심기술을 개발하고 이를 다양한 시장으로 유도하고 있기 때문이다. 이와 더불어 효과적인 팀워크, 모니터링 가능성 및 확장성을 위한 ML 프로세스 표준화가 MLOps에 적용되면 향후 산업에 더욱 채택되어 시장확장이 예상되고 있다.

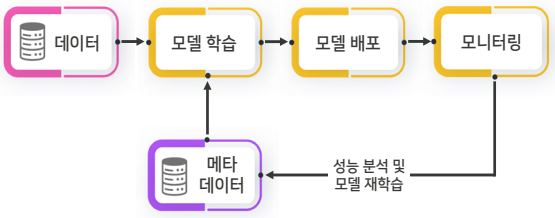
앞서 언급한 바와 같이 MLOps 서비스의 산업 적용이 확대됨에 따라 관련 기술에 대한 관심이 높아지고 있는 상황이다. 따라서, MLOps를 보다 쉽게 접근하기 위해서 먼저 핵심 모듈인 공통 기본 요소들을 소개하고 주요 MLOps 서비스를 제공하는 글로벌 기업들 솔루션 및 공개 프로젝트에 대한 개발 환경을 서술하고자 한다.

1. 머신러닝 워크플로우

머신러닝 워크플로우는 그림 1과 같이 데이터 전처리 및 탐색적 데이터 분석, 데이터 변환 및 생성, 하이퍼파라미터 최적화, 신경망 학습, 신경망 배포의 실행 단계로 구성된다. 각 모듈에서의 역할은 다음과 같다.

데이터 전처리 및 탐색적 데이터 분석에서 신경망 학습을 하기 위해서는 데이터 세트가 필요하며 이를 위한 처리이다. 데이터 변환 및 생성에서는 신경망 학습을 위한 데이터 크기 및 증강을 의미하며, 하이퍼파라미터 최적화에서는 최적의 학습을 수행하기 위해서는 다양한 파라미터의 초기화 값 설정이 중요한데 이를 찾고 설정하는 과정이다. 신경망 학습과 배포는 주어진 데이터를 바탕으로 최적의 신경을 생성하고 생성된 신경망을 디바이스에 탑재하는 과정이다. 이전에는 이런 과정을 하나하나 수동적인 단계를 밟아 수행되었는데, 이를 자동화한 것이 머신러닝 워크플로우 자동화이다. 이 기능은 MLOps에서 중요한 핵심요소이다.

산업현장에서는 예전과는 달리 한번 개발된 신경망이 제품의 수명과 함께하는 것이 아니라, 데이터의 지속적 발생에 대응하고 이에 대한 성능을 개선



출처 게티이미지뱅크, 무단 전재 및 재배포 금지

그림 3 지속적 신경망 생성의 기본 구조

하기 위해 지속적인 학습을 수행하고 이를 관리하는 모듈들이 추가되어야 한다. 즉, 신경망 모델을 디바이스에 배포하고 주기적으로 유입되는 데이터를 바탕으로 신경망 학습을 수행한 후 평가를 통해서 업데이트의 여부를 결정한다. 이를 자동적이면서 지속적으로 처리하는 것이 MLOps의 기본 개념이다.

그림 3은 머신러닝 워크플로우에 모니터링 기능과 모델 재학습 및 관리 모듈이 포함된 기본 개념도이다. 현재 서비스되고 있는 주요 MLOps들은 이런 요소기술을 포함하고 있으며, 추가적인 기능들을 제공한다.

대표적인 MLOps 도구로는 퍼블릭 클라우드에는 Google Vertex AI[2], MS Machine Learning Azure[3], Amazon SageMaker[4] 등이 있으며, 공개 소스코드 기반 프로젝트로는 Kubeflow가 유명하다. 다음은 각 자사의 클라우드에 신경망 생성 및 배포를 지원하는 MLOps의 공통적인 요소들을 제시하고, 오픈 프로젝트인 Kubeflow에 대해 보다 자세히 알아본다.

2. 글로벌 상용도구 MLOps

자사의 클라우드 서비스를 위해서 각자의 특색있는 MLOps를 개발자에게 지원하고 있으나, 개발자의 편의성 및 개발 환경 제공을 위해 공통적인 기능들을 살펴보면 다음과 같다.

가. 머신러닝 워크플로우

머신러닝 워크플로우는 앞장에서 서술한 바와 같이 AutoML을 위한 핵심 요소들이며, 이를 이용한 부가적 기능 추가로 MLOps의 활용성을 강화한다. 각 핵심요소들의 파이프라인 수행으로 신경망이 생성되고 디바이스에 배포된다.

나. 신경망 모델 관리 작업 및 자동화

MLOps에서 다양한 환경에서 프로젝트를 생성하여 실행하면서 결과물로 나오는 수천 개의 신경망 모델 배포 및 관리의 간소화 기능을 포함하고 있다. 이는 전주기 및 실시간 예측을 위한 엔드투엔드(End-to-End) 포인트 기반으로 관리하여, 모델을 더 빠르게 배포하고 성능을 점검한다. 이러한 과정을 통해 반복 가능한 파이프라인을 수행하여 지속적 통합 및 지속적 배포(CI/CD: Continual Integration/Continual Deployment)를 위한 워크플로우를 자동화한다. 또한, 레지스트리 및 관리형 기능 저장소를 사용할 수 있으며, 작업 영역 간 공동 작업을 위해 다수 개발자가 머신러닝 신경망 모델을 공유하고 검색할 수 있다. 모델 성능 메트릭을 지속적으로 모니터링하고, 데이터 드리프트를 감지하고, 필요시 재학습 트리거를 통해 모델 성능을 개선한다. 트리거는 일정 시간 또는 이벤트에 대한 응답에 따라 자동으로 실행되며, 이를 바탕으로 신경망 모델 생성의 자동화 기능이 완성된다. 이 단계의 결과물은 신경망 모델 레지스트리로 푸시되고 통합 관리된다.

다. 개발자 요구에 맞는 모델 선택

다양한 API 기반으로 공개 소스코드 기반 신경망 모델뿐만 아니라 개발자의 개발 모델을 한 저장 공간에 저장함으로써 필요한 신경망 모델을 액세스하여 원하는 ML 프로젝트를 빠르게 생성 가능한 기능을 제공한다. 따라서, 개발자는 모델을 직접 가져와

서 사용하거나, 생성형 AI 도구를 통해 모델을 미세 조정(Fine Tunning)하거나, 편집 노트북으로 모델을 배포하는 등 다양한 작업을 수행할 수 있다.

라. 로우 코드 및 노코드 도구

다양한 전문성을 갖춘 개발자뿐만 아니라 머신러닝에 초보자도 머신러닝 워크로드를 사용할 수 있도록 로우 코드/노코드 도구와 GUI 기반 간편 인터페이스 기능을 제공한다. 여기에 생성형 AI 도구를 사용하면 개발자가 간단한 인터페이스를 통해 원하는 환경에 맞게 신경망 모델을 미세 조정하고 배포할 수 있다. 또한, API를 통해 개발자는 선행 학습된 신경망 모델을 쉽게 호출하여 원하는 신경망 모델을 신속하게 생성할 수 있다.

마. 학습 신경망과 배포 신경망의 관리 모듈

MLOps 프레임워크는 다수의 개발자를 위한 개발 서비스로, 기존 개발자에 의해 생성된 신경망 모델을 저장하고 관리하고 있으며, 디바이스나 응용 및 데이터 셋에 대해서는 다른 개발자가 재사용 가능하도록 지원한다. 기존의 학습 및 배포를 통해 동일 개발 환경에서는 빠르게 적용할 수 있다. 따라서, 이런 상황을 고려하여 관련 데이터를 저장하고 재사용할 수 있는 환경을 마련하는 기능을 제공한다. 또 다른 예로 한번 구동되었던 소스 코드는 소스 저장소로 푸시되고 재사용 가능하다.

바. 지속적 통합(CI) 및 배포(CD)

신경망을 생성하고 배포하는 개발 과정에서 소스 코드를 빌드하고 다양한 테스트를 실행한다. 테스트에서 품질이 결정되면, 관련 소스 코드에 대해서 버전 관리를 하게 되며, 이후 단계에서는 배포될 파이프라인 구성요소들(패키지, 실행 파일, 신경망 모델 등)을 관리한다.

지속적 통합 단계에서 생성된 신경망 모델은 개발자가 원하는 디바이스에 배포한다. 이후 단계에서는 배포된 신경망 모델을 다음 개발자가 재사용할 수 있도록 개발환경과 신경망 모델 및 성능 평가 등이 관리된다. 성능 평가는 재학습이 수행된 이후 지속적 배포를 수행할지 결정할 중요한 요인이 된다.

Git와 연동 기능을 제공하여, 사용자가 개발한 신경망 모델을 MLOps 내에 학습 파이프라인에 적용할 수 있으며, 배포를 위한 추론 코드도 Git와 연동할 수 있게 지원한다.

지속적 통합과 지속적 배포는 머신러닝 자동화 관리 요소와 더불어 산업환경에서 MLOps 적용에 주요 기능을 담당하게 된다.

사. 모니터링

머신러닝 워크플로우가 실행되는 경우, 실시간 데이터를 기반으로 신경망 모델 성능을 수집하고 통계 분석을 수행한다. 분석된 데이터를 통해 이 단계에서는 파이프라인을 실행을 관리하거나 새 신경망 모델 개발 주기를 관리하며, 트리거를 구동하기도 한다.

3. 공개 프로젝트 Kubeflow

Kubeflow는 공개 소스코드 기반 엔드투엔드 AI 신경망 프레임워크이다. 머신러닝 워크플로우의 신경망 모델 학습부터 신경망 모델 배포 단계까지 모든 작업에 필요한 도구와 환경을 쿠버네티스(Kubernetes) 위에서 동작하도록 Kubeflow 컴포넌트로 제공한다. 앞서 언급한 상용 MLOps와는 달리 공개 소스코드로 개발자에게 제공되고 있어, 클라우드 서비스 선택에 자유도가 있으며 기능을 추가하는 데 이점이 있다. 주요 구성요소는 Kubeflow 컴포넌트와 Kubeflow 외부 추가 기능으로 구성되어 있으며 각

각의 기능 및 특징으로는 다음과 같다.

가. 중앙 대시보드

중앙 대시보드(Central Dashboard)는 Kubeflow 및 관련 생태계 구성요소에 대해 인증된 웹 인터페이스를 제공한다. 클러스터에서 실행되는 구성요소의 UI를 표출하여 머신러닝 플랫폼 및 도구의 허브 역할을 한다. 주요 기능은 프로필 및 네임스페이스를 기반으로 한 인증 및 권한을 부여하고, Kubeflow 구성요소의 사용자 인터페이스에 액세스를 지원하며, 타사 애플리케이션에 대한 링크를 사용자 정의하고 포함하는 기능을 포함한다.

나. Kubeflow Notebooks

Kubeflow Notebooks는 Kubernetes 클러스터 내에서 웹 기반 개발 환경을 실행하는 방법을 제공한다. JupyterLab, RStudio 및 Visual Studio Code(코드 서버)에 대한 기본 기능을 지원하며, 개발자는 워크스테이션에서 로컬이 아닌 클러스터에서 직접 노트북 컨테이너를 생성할 수 있으며, 관리자는 필수 패키지가 사전 설치된 조직에 표준 노트북 이미지를 제공할 수 있다. 액세스 제어는 Kubeflow의 RBAC(Role-Based Access Control)로 관리되므로 조직 전체에서 더 쉽게 노트북을 공유할 수 있도록 하였다.

다. 학습 운용

학습 운용은 PyTorch, TensorFlow, XGBoost 등과 같은 다양한 ML 프레임워크를 사용하여 생성된 머신러닝 모델의 미세 조정 및 확장 가능한 분산 교육을 위한 Kubernetes 기반 프로젝트이다.

개발자는 HuggingFace, DeepSpeed 또는 Megatron-LM과 같은 다른 ML 라이브러리를 학습 운용과 통합하여 Kubernetes에서 ML 학습을 조정할 수 있다.

또한, Kubernetes 워크로드를 사용하여 Kubernetes Custom Resources API 또는 Python SDK를 통해 대규모 모델을 효과적으로 학습할 수 있다.

라. Katib(AutoML)

AutoML은 머신러닝 모델의 예측 정확도와 성능을 높이기 위한 반복 실험을 자동화하는 도구이다. Kubeflow에서는 카티브(Katib)를 사용하여 신경망 자동생성 기능을 제공한다. Katib는 하이퍼파라미터 튜닝(Hyper Parameter Tuning), 신경망 구조 탐색(NAS: Neural Architecture Search), 신경망 탐색 조기 중단 등의 기능을 제공한다. 하이퍼파라미터 최적화는 모델의 하이퍼파라미터를 최적화하는 작업이고 NAS는 신경망 모델의 구조, 노드 가중치 등 신경망 아키텍처를 최적화하는 작업으로 다양한 머신러닝 최적화 알고리즘을 지원한다.

마. Kubeflow Pipelines(KFP)

Kubeflow Pipelines(KFP)은 머신러닝 워크플로우를 구축하고 도커 컨테이너를 사용하여 이식 가능하고 확장 가능한 신경망 모델 생성을 위한 머신러닝 워크플로우 자동화 도구이다.

KFP를 사용하면 KFP Python SDK를 사용하여 구성요소와 파이프라인을 작성하고, 파이프라인을 중간 표현 YAML로 컴파일하고, 파이프라인을 제출하여 공개 소스코드 기반 KFP 백엔드 또는 다른 머신러닝(예: 구글 Vertex AI) 파이프라인과 같은 KFP 호환 백엔드에서 실행이 가능하다. KFP를 사용하게 되면, 기본적으로 엔드투엔드 워크플로우를 구성할 수 있으며, 맞춤형 머신러닝 구성요소를 생성하고 기존의 구성요소로 생성된 워크플로우를 활용 가능하다. 또한, 파이프라인을 정의하고 자동 실행함으로써 신경망 생성 및 관리를 쉽게 할 수 있으며, 추적 및 시각화도 가능하다. 플랫폼 중립적인

YAML(YAML Ain't Markup Language)을 이용한 파이프라인 정의를 통해 플랫폼 간 사용이 가능하도록 지원한다.

바. KServe

KServe는 쿠버네티스 환경에서 서버리스 추론을 활성화하고, TensorFlow, XGBoost, scikit-learn, PyTorch 및 ONNX와 같은 일반적인 머신러닝 프레임워크에 대한 고성능 및 높은 추상화 인터페이스를 제공하여 머신러닝 워크플로우가 실행될 수 있도록 지원한다. 즉, 임의의 프레임워크에서 ML 모델을 제공하기 위한 Kubernetes 사용자 정의 리소스 정의를 제공한다. 또한 자동 크기 조정, 네트워킹, 상태 확인 및 서버 구성의 복잡성을 캡슐화하여 GPU 자동 크기 조정, Zero Scaling, ML 배포에 카나리아 롤아웃과 같은 서비스 기능을 제공한다. 예측, 사전 처리, 사후 처리 및 설명 가능성을 제공함으로써 머신러닝 추론 서버에 대해 간단하고 플러그인 가능하며 완벽한 계층을 구성하도록 지원한다.

2절과 3절에서 언급된 주요 4가지 MLOps에 대한 지원 기능을 정리해 보면 표 1과 같이 요약할 수 있다.

Ⅲ. 타겟 지향형 신경망 생성 프레임워크 (TANGO)

기존의 MLOps는 신경망 생성을 하고 신경망 배포를 각각 독립적으로 수행하여 디바이스에 최적화된 신경망을 만드는 데 애로점이 있었다. 이 문제점을 해결하기 위해서 디바이스 성능을 고려하여 요구되는 신경망을 생성한 후 디바이스의 구동 환경(OS 및 Pytorch, TensorRT, TVM 등)에 맞게 다시 한번 최적화 프로세스를 수행하여 디바이스 최적 신경망을 생성하고 구동할 수 있는 프레임워크[6,7]를 제안하였다.

그림 4는 기존 MLOps의 요소 및 기능을 가지면서, 추가적으로 보완되어야 할 모듈에 대해서 하이라이트로 표시하였다. TANGO(Target Aware No-code Neural network Generation and Operation framework) 주

표 1 MLOps별 기능 비교

	Google Vertex AI[2]	MS Azure ML[3]	AWS SageMaker[4]	Kubeflow[5]
구동 플랫폼	• GCP 기반 Vertex AI 플랫폼	• 애저 기반 MLOps 플랫폼	• AWS 기반 MLOps 플랫폼	• Kubernetes 기반 공개 소스
데이터 준비 작업	• 데이터 레이블링 서비스 지원 • 데이터 셋 관리 지원	• 데이터 레이블링 서비스 지원	• 데이터 레이블링을 위한 Ground Truth 지원	• 주피터 노트북 이용
신경망 모델 학습 /병렬 처리	• 통합 메타 데이터 기반 파이프라인 실행 • 병렬처리 지원	• 병렬처리를 위한 파이프라인 기반 학습	• 파이프라인 지원	• 파이프라인 지원 • 하이퍼파라미터 튜닝 지원 AutoML
신경망 모델 배포	• 커스텀 기반 배포	• 커스텀 기반 배포	• 커스텀 기반 배포	• KFServing • 전/후처리 추론 단계 지원
병렬화, 저지연, 멀티모델 지원	• 트래픽 분산 • 저지연 지원	• GPU와 트래픽 분산처리	• 트래픽 분산 • 멀티모델 지원 • 저지연 위한 탄력적 추론 지원	• 자동 스케일링 • 트래픽 분산 • 멀티모델 서비스 • GPU 통합 실행
신경망 모델 성능 및 모니터링	• 지연 시간 모니터링	• 지연 및 HW 자원 모니터링	• 커스텀 모니터링 스케줄 지원 • CloudWatch	• 자원 및 High-level 매트릭스 지원
신경망 관리	–	• 작업공간 관리 지원	• 버전, 그룹, 기관별 신경망 관리 지원	–

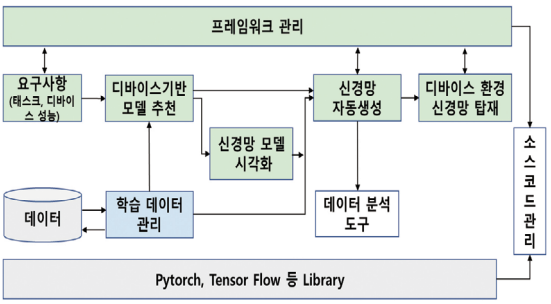


그림 4 TANGO[6] 프레임워크 구조

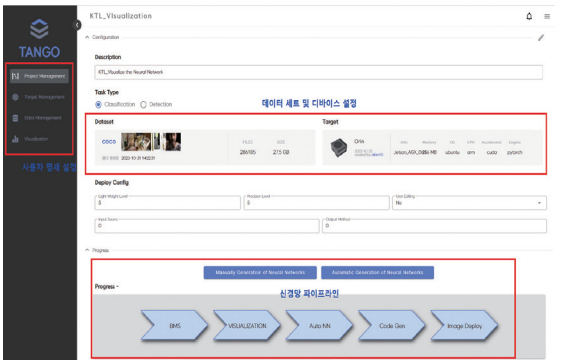


그림 5 TANGO 사용자 인터페이스

요 모듈의 구체적인 특징 및 기능에 대해서 서술하면 다음과 같다.

1. 프레임워크 구조 및 관리

TANGO는 Docker 기반 MSA(Micro Service Architecture)로 각 모듈이 연동되며, 각 모듈은 파이프라인 구조로 실행되도록 설계되었다. 즉 입력 처리 모듈, 신경망 모델 추천, 신경망 생성 및 신경망 배포의 구성요소들을 각각의 프로세스로 설정하고 상황에 맞게 순차적으로 처리할 수 있도록 고안하였다. MSA 구조의 장점에 맞게 각 모듈은 독립적으로 구동 및 디버깅이 가능하며, 기존 모듈에 대체 가능한 모듈과 교체가 가능하도록 하였다. 그리고 향후에 필요에 의해서 구성요소들을 추가할 경우에 다른 모듈의 구동에 영향이 미치지 않도록 하였으며, 다양한 모델을 기반으로 하는 신경망 생성의 경우 같은 위치에 동일 기능이 가능한 다양한 신경망 모델을 배치할 수 있도록 구조를 제안하였다.

2. UI 기반 입력 정보 명세

TANGO 프레임워크는 사용자의 간단한 입력 사항을 기반으로 프로젝트를 생성하고, 관련 입력 정보를 다음 파이프라인 구성모듈로 자동 전달하여 구동되도록 하였다. 신경망 생성을 위해서는 기본적으로 요구되는 정보가 있는데, 신경망의 종류를 결정하는 Task 타입(예: 이미지 분류 vs 영상 객체 탐지)에 대한 정보와 디바이스의 연산 처리속도 성능이다. 신경망 배포를 위해서는 탑재 디바이스에서의 HW 및 SW 구동 환경이 제공되도록 입력 UI를 구현하였다. 그림 5는 입력 정보를 입력하는 것으로, 데이터 타입 및 디바이스의 사양을 결정하는 것을 보여주는데, 타겟 디바이스 설정에서 SW 환경을 설정 및 탑재를 위한 고급 옵션들이 포함되어 있다.

3. 신경망 생성 모듈

신경망 모델은 제공하고자 하는 응용(Task)과 사용하는 데이터에 따라 다양하게 변경될 수 있다. TANGO에서는 산업분야에서 많이 적용되는 시각 데이터에 관련된 이미지 분류와 객체 탐지를 지원하도록 구현되었다.

이미지 분류의 경우, 현재 SOTA(State-Of-The ART) 모델보다는 산업환경에서 기본적으로 많이 사용되는 Resnet과 Densenet 기반으로 설계되었으며, 디바이스 성능에 따른 추론 시간을 고려하여 5개의 복잡도를 제공하도록 하였다. 앞서 언급했듯이, 프레임워크 구조가 MSA 방식으로 되어 있어 향후에

응용 및 데이터에 맞게 다른 신경망 모델을 사용할 수 있다.

영상의 객체 탐지의 경우에는 정확도뿐만 아니라 적용 디바이스의 실시간 추론 시간을 고려하여 one-stage로 처리하는 Yolo 모델[8]을 기반으로 하여 구현하였다. 신경망 생성을 위한 입력 파일로 YAML 파일을 받아서 학습을 수행하되 3가지 방식으로 기능을 확장하였다. 첫째, 기존 Yolo 모델보다 성능 향상을 위해서 Neck 구조에서 피쳐의 크기를 재조절하고 3D 확장하는 알고리즘을 추가(Bag of Specials)하여 정확도를 높였다. 둘째, 한 번의 학습으로 다양한 디바이스에 적용 가능한 Supernet 기반 NAS(Neural Architecture Search) 구조[9]로 설계되어, 요즘 많이 사용되는 스마트폰용 디바이스에 적용 가능하도록 하였다. 셋째, 모델별로 최적의 학습 파라미터가 존재하는데, TANGO에서는 하이퍼파라미터를 최적화하여 학습의 속도와 정확도를 높였다.

4. 데이터 레이블링

지속적인 데이터 유입에 따른 학습이 원활하게 수행되려면, 데이터 전처리 및 분류가 자동적으로 이루어져야 한다. 이를 위해서 능동적 학습 기반 데이터 레이블링[10] 기능이 추가되고 있으며, 이로써 지속적 통합/지속적 배포를 위한 지속적 학습(Continual Training) 기능이 제공된다.

Ⅳ. 결론

다양한 산업 분야에서 업무 효율화를 위해서는 인공지능 신경망 응용 개발의 필수요소가 되어 가고 있다. 이런 상황에서 클라우드 서비스를 갖춘 글로벌 기업에서는 신경망 개발을 용이하게 할 수 있는 MLOps를 제공하고 있으며, 이들의 공통적인 기능들을 살펴보았다. 글로벌 기업들의 자사 위주의 클라우드 서비스 사용을 제한하는 것과는 달리 Kubeflow는 공개 프로젝트를 지향하여 다양한 시스템에 적용할 수 있도록 기능을 갖추었다. 한편, 타겟 적응형 기반인 TANGO도 개발 중이며 공개로 진행하고 있음을 살펴보았다.

용어해설

AutoML 산업현장의 실제 문제를 해결하기 위해 적용하는 머신러닝을 자동화하는 프로세스로, 원시 데이터부터 데이터 처리, 신경망 생성 및 배포까지 포함

MLOps 머신러닝(ML)과 모델 운용(Operations)을 합친 용어로, 프로덕션 환경에서 머신러닝 모델을 안정적이고 효율적으로 배포 및 유지 관리하는 것

약어 정리

CI/CD	Continual Integration/Continual Deployment
KFP	Kubeflow Pipelines
ML	Machine Learning
MLOps	Machine Learning Operations
MSA	Micro-Service Architecture
NAS	Neural Architecture Search
SOTA	State-Of-The ART
YAML	YAML Ain't Markup Language

참고문헌

[1] Market and Market, "MLOps Market by Component (Platform and Services), Deployment Mode (Cloud and On-premises), Organization Size (Large Enterprises and SMEs), Vertical (BFSI, Healthcare and Life Sciences, Retail and eCommerce, Telecom) and Region – Global Forecast to 2027," 2022, https://chosareport-korea.com/mnmtc8518/

[2] Google, "Google Vertex AI," https://cloud.google.com/vertex-ai?hl=ko

[3] Microsoft, "MS Azure Machine Learning," https://azure.microsoft.com

[4] Amazon, "Amazon SageMaker," https://aws.amazon.com/pm/sagemaker

[5] Kubeflow Homepage, https://www.kubeflow.org/

[6] ETRI, "TANGO Project," GitHub, https://github.com/ML-TANGO/TANGO

[7] 김선태, 조창식, "디바이스 적응형 신경망 생성 및 배포 구현," 전자공학회논문지, 제61권 제1호, 2024, pp. 27-33.

[8] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object Detectors," arXive preprint, Jul. 2022, https://doi.org/10.48550/arXiv.2207.02696

[9] I. Shin, C. Cho, and S.-T. Kim, "Method for Expanding Search Space With Hybrid Operations in DynamicNAS," IEEE Access, vol. 12, 2024, pp. 10242-10253.

[10] J. Auh, C. Cho, and S.-T. Kim, "Improved contrastive learning model via identification of false-negatives in self-supervised learning," ETRI J., online published, 2024, https://doi.org/10.4218/etrij.2023-0285