






TANGO 커뮤니티 소개

성명 조창식

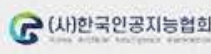
소속 ETRI

SUBJECT

인공지능 기술의 대중화 (AI Democratization)를 위한
TANGO 커뮤니티 3회 컨퍼런스

주관 ETRI () 주최  과학기술정보통신부  정보통신기획평가원

후원



목 차

1

TANGO 프로젝트

3

1. Intro
2. 모든 개발 과정을 Github에서
3. 신경망 응용 개발의 어려움
4. 해외 MLOps 동향

2

TANGO 차별성

13

1. Detection 자동 생성/학습
2. 다양한 배포 환경 지원
3. 통합 파이프라인 지원
4. Well Defined SW architecture

3

성과 확산

22

1. 실증을 통한 탱고 검증
2. 기술이전 및 사업화
3. Future Work

Intro

Target Aware No-code neural network Generation and Operations framework

(**타겟 인지형 No-code 기반 신경망 자동 생성/배포 통합개발 프레임워크**)

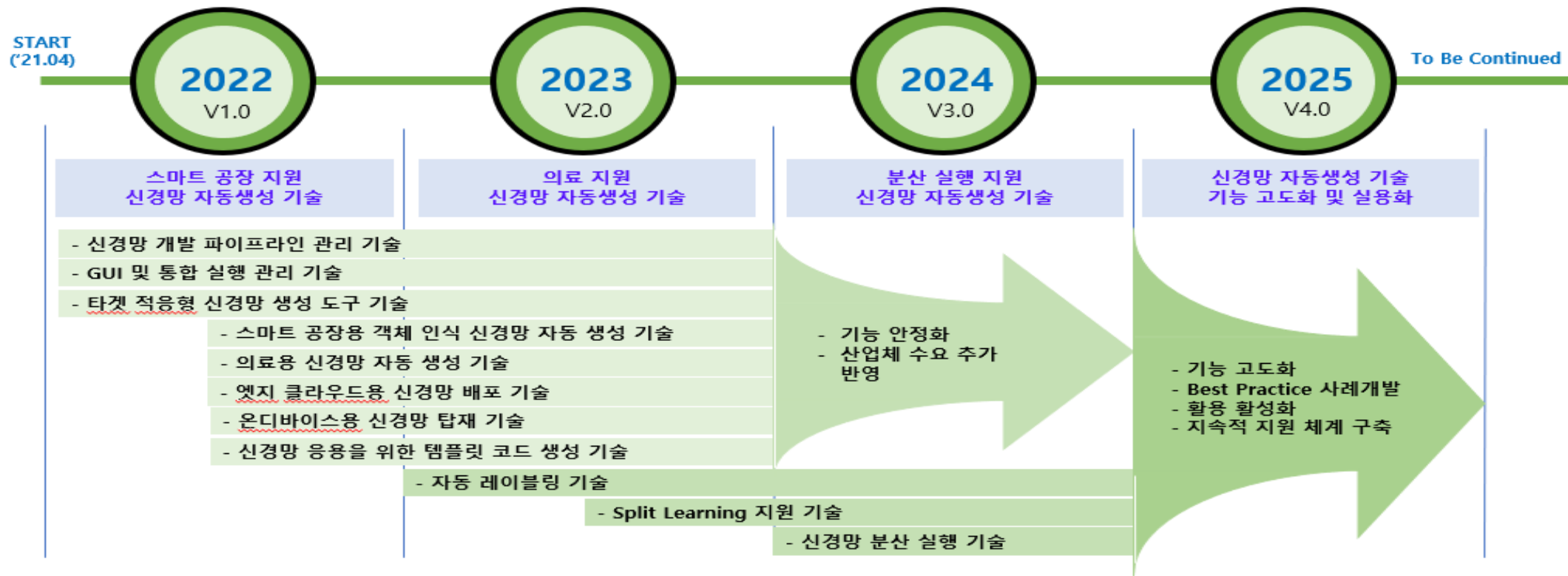
타겟 장비(클라우드, 엣지, 온디바이스)의 HW 성능 특성을 인지하여 신경망을
잘 모르는 산업현장(공장, 의료) 사용자도
최적의 신경망을 자동생성/배포할 수 있도록 지원하는 통합개발 프레임워크



AI 대중화 시대의 필수 전략 기술
(AI Democratization)

모든 개발 과정을 Github에서

매년 두 번의 릴리즈 버전을 완성하고 하반기에 공개SW 세미나 추진



모든 개발 과정을 Github에서

main 브랜치는 항상 컴파일/실행 버전 유지, 문서는 위키 페이지로 단일화

소스 (브랜치) 관리

- 담당자별 브랜치에서 개발
- 담당자별 단위 테스트
- 담당자별 통합 테스트
- 통합 테스트(수동),
No CI/CD yet
- main 브랜치에 merge/push

문서 관리

- 모든 문서는 위키 페이지에 공유
 - 단일 참조 포인트
- 개발 가이드
 - TANGO Architecture
 - YAML (컨테이너 통신)
 - Rest API
 - Container Port Map

이슈 관리

- 이슈 관리
 - GitHub Issues
- Backlogs 관리
 - GitHub Project,
Kanban 스타일

<https://github.com/ML-TANGO/TANGO/wiki>

TANGO 히스토리

- 2021.04 신경망 자동생성 통합개발 프레임워크 과제 시작
- 2022.02 비공개 GitHub 저장소 운영
- 2022.09 공개 GitHub 저장소 운영
- 2022.10.31 Pre 릴리즈 (tango-22.11-pre1)
- **2022.11.01 1회 TANGO 커뮤니티 컨퍼런스 (AT센터 세계로룸)**
94개 기관 158명이 참석
- 2022.11.30 22년 하반기 정식 릴리즈 (tango-22.11)
- 2023.05.31 23년 상반기 정식 릴리즈 (tango-23.06)
- 2023.10.23 23년 하반기 정식 릴리즈 (tango-23.10)
- **2023.11.01 2회 TANGO 커뮤니티 컨퍼런스 (과학기술회관)**
96개 기관 215명이 참석
- 2024.05.31 24년 상반기 정식 릴리즈 (tango-24.05)
- 2024.11.21 24년 하반기 정식 릴리즈 (tango-24.11)
- **2024.12.05 3회 TANGO 커뮤니티 컨퍼런스 (과학기술회관)**



신경망 응용 개발의 어려움

데이터 준비

신경망 학습

신경망 배포

응용SW 개발



데이터
가공



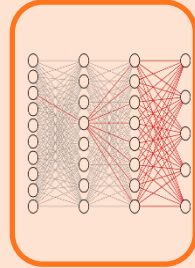
라벨링



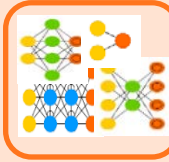
도메인 전문가



코딩



학습



신경망
모델



ML scientist

클라우드 배포



온디바이스
가속



엣지
컴퓨팅



ML engineer

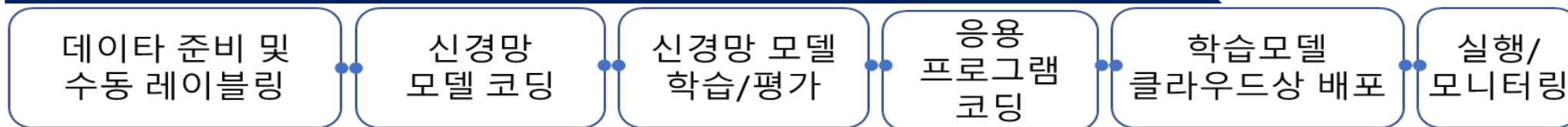


도메인 전문가

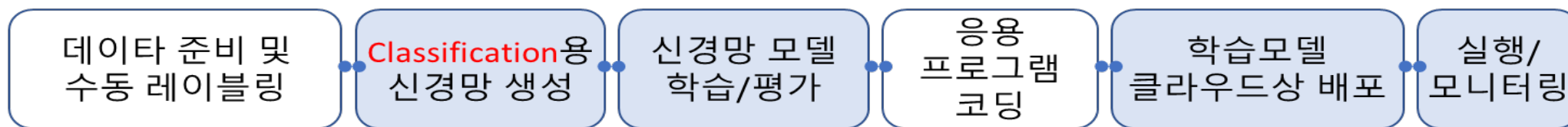
도메인 지식과 신경망 전문지식에 대한 고도의 개발경험 요구

신경망 응용 개발의 어려움

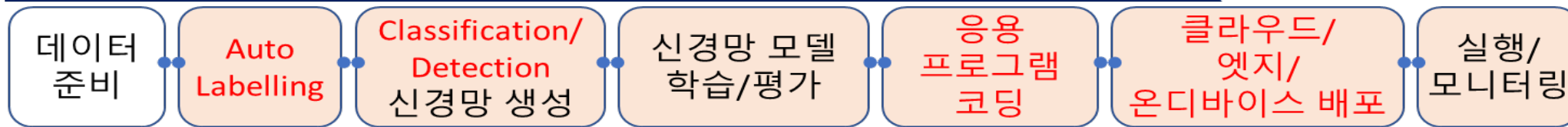
클라우드 기반 신경망 개발: 수동 코딩



기존 AutoML: Classification 중심, 부분 자동화



목표 기술: Detection 지원, No Code 및 다종 디바이스 지원



해외 MLOps 동향

퍼블릭 클라우드 AutoML 기반 MLOps 도구 각축

구글 Vertex.AI



MS AzureML



Amazon SageMaker



오픈소스 Kubeflow



수동 프로그래밍을 쉽게 하기 위해, 라이브러리/API를 추상화하는 방향으로 진화
AutoML API 제공, Python 라이브러리를 사용하여 쉬운 코딩 지향

중급 이상의 신경망 전문지식 요구

해외 MLOps 동향

퍼블릭 클라우드 AutoML 기반 MLOps 도구 각축

MLOps 도구 특징

- 다양한 인공지능 응용 지원
- 다양한 AutoML(NAS, HPO) 알고리즘 지원
- 다양한 배포환경 지원 (클라우드, 엣지, 온디바이스)
- 손쉬운 웹 UI 제공

지원 응용

- Image Classification
- Tabular Classification
- Tabular Regression
- Text Classification
- Object Detection
- Text Embedding
- Question Answering
- Sentence Pair Classification
- Image Embedding
- Named Entity Recognition
- Instance Segmentation
- Text Generation
- Text Summarization
- Semantic Segmentation
- Machine Translation

지원 알고리즘

- ENAS
- DARTS
- P-DARTS
- SPOS
- CDARTS
- ProxylessNAS
- ...

주로, Tabular 데이터에 대한 AutoML 적용[ML],
이미지의 경우 Classification에 집중, 하이퍼파라미터 최적화 위주[DL]
배포는 자사 클라우드에 최적화

해외 MLOps 동향

Classification, Detection, Segmentation 비교



단순히 단일
이미지 분류



의료에 적용



여러 객체 분류와
위치까지 표시



공장에 적용
(TANGO의 주요 타겟)



객체의 윤곽까지
표시



데이터 라벨링에
장시간 소요

해외 MLOps 동향

Classification, Detection, Segmentation 산업체 적용 예

Classification (폐결핵검사)



○ 폐질환 분석

- 정상과 폐결핵인지 분류
- 폐결핵은 5개 병으로 세분화
- 영상의학과 의사가 라벨링
- 특징벡터 기반 연합학습
- Densenet 백본 사용

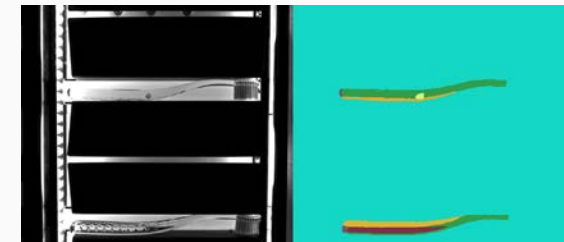
Detection (용접불량 검사)



○ 파이프 성형시 용접불량 탐지

- 파이프 X선 촬영 비파괴검사
- 용접부위의 정확한 불량부위 검출 (과다용접, 기공, 크랙 부위 등)
- 파이프 X선 이미지 영상의 육안 검사로 라벨링
- YOLO 신경망 사용

Segmentation (치솔불량 검사)



○ 치솔 불량부위 검출

- 정상 치솔과 불량 치솔을 구분
- 대표적인 불량은 손잡이 기포
- 일반인도 육안으로 라벨링 가능
- 라벨링 소요 시간 많음
- Unet 신경망 적용

II. TANGO 차별성

Detection 자동생성/학습

Detection을 지원하는 신경망 자동생성 도구(세계최고 성능 추구)

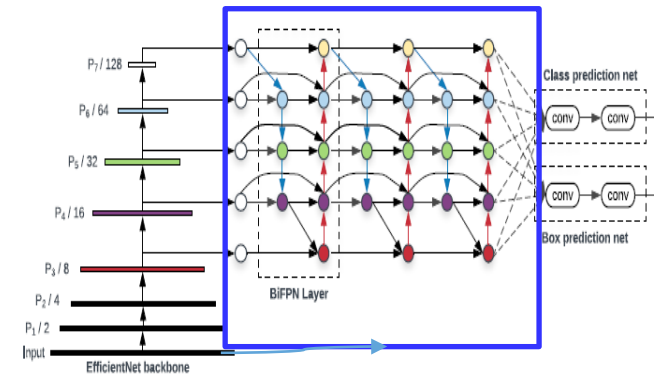
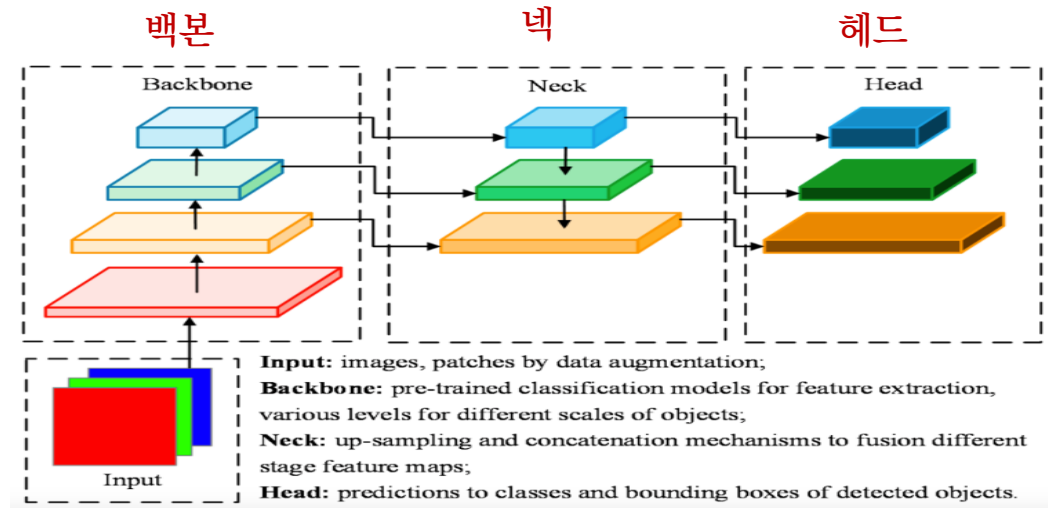


Figure 3: EfficientDet architecture - It employs EfficientNet [39] as the backbone network, BiFPN as the feature network, and shared class/box prediction network. Both BiFPN layers and class/box net layers are repeated multiple times based on different resource constraints as shown in Table 1.

- Classification는 백본만 있음 (Resnet, Densenet)
- Object Detection(객체 탐지)는 백본, 넥, 헤드로 구성됨
- Detection 분야는 아직도 진화 중 (YOLO3/4/5/6/7/8/9/10/11, PPYOLO, YOLOX, ScaledYOLO ,,,)

백본, 넥 계층 탐색을 통하여 정확도, 성능을 고려한 신경망 자동생성

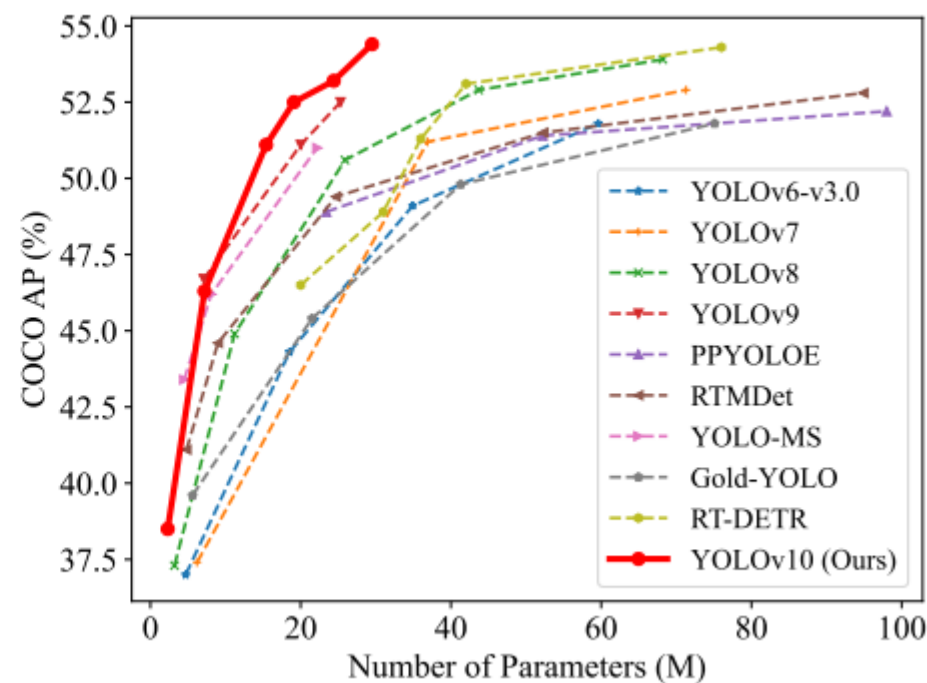
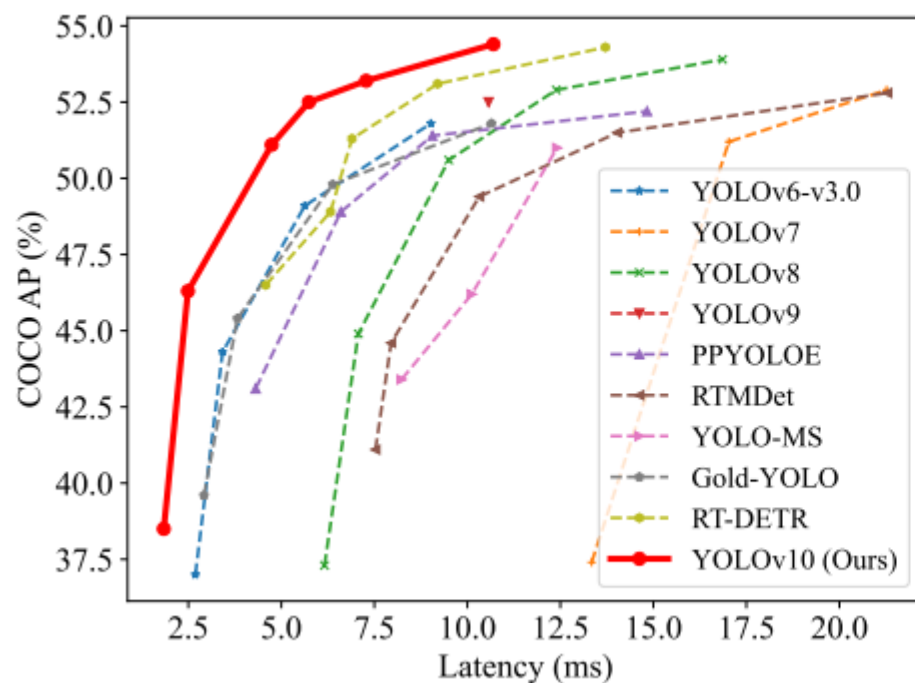
II. TANGO 차별성

Detection 자동생성/학습 (Yolo History)

	저자	기관	날짜	백본	넥	헤드	mAP(정확도)
yolo1	Joseph Redmon	Washington 대학	2016년	DarkNet : GoogLeNet 변형 (Inception 대신 1x1 conv 사용; 24... x convolution)	-	2 x FC	63.4 mAP (COCO2007)
yolo2	Joseph Redmon	Washington 대학	2017년	DarkNet19	2 scale concatenation 적용	FC를 1x1 Conv로 대체 (완전한 CNN 구성)	69.2 mAP, recall 88% (COCO2007+2012)
yolo3	Joseph Redmon	Washington 대학	2018년	Darknet53 (Bottleneck 구조 적용)	FPN	3개의 Scale을 사용	이후 COCO2012로 통일
yolo4	Alexey Bochkovskiy	Academia Sinica	2020년4월	CSPDarkNet53 (Darknet53 구조를 유지하면서 Bottleneck에 CSP를 적용)	SPP, PAN (CSP 미적용)	YOLOv3 head 그대로 사용	COCO mAP 43.5%, 65.7 FPS
yolo5	Glenn Jocher	Ultralytics	2020년6월	CSPDarkNet53의 변형 (CSPBottleneck을 적용한 반복 횟수와 순서가 다름) CSP에서 convolution 3개 사용한 버전(C3) 이용	SPP, PAN (CSP 적용)	YOLOv3 head의 변형	yolo5 s/m/l/x
yolo6	Chui Li	Meituan Inc	2022년7월	EfficientRep	CSPSPPF, BiPAN (Rep 적용) (CSP 적용)	Efficient decoupled head (YOLOX)	yolo6 n/s/m/l/16
yolo7	Chien-Yao Wang	Academia Sinica	2022년9월	E-ELAN Rep 적용	PAN, Rep 적용	큰 모델의 경우 Auxiliary Head	yolo7 X/W6/E6/D6/E6E
yolo8	Glenn Jocher	Ultralytics	2023년1월	YOLOv5에서 일부 수정 (C2f: Convolution 2개인 CSP 버전)	SPPF, PAN	Decoupled head	yolo n/m/x YOLOv8m 50.2의 mAP
yolo9	Chien-Yao Wang	Academia Sinica	2024년2월	GELAN (E-ELAN의 일반화, Rep 적용, CSP 사용)	SPP, PAN, (PGI Fusing 연산 도입, Rep 적용)	Dual / Triple Head (학습시)	yolo9 s/m/c/e
yolo10	Ao Wang	Tsinghua 대학	2024년5월	enhanced version of CSPNet	PAN	One-to-Many Head, One-to-One Head	yolo10 n/s/m/b/x yolo10x: 54.4% mAP _{val50-95}
yolo11	Glenn Jocher	Ultralytics	2024년9월	C3k2 (convolution 3개 사용한 CSP의 더 빠른 속도를 구현한 버전)	PAN	Single Head	yolo11x: 54.7% mAP _{val50-95} on COCO
RT-DETR	Xian Zhao	Baidu	2023년4월 CVPR 2024	기존 백본 사용(HGNetV2 large/extra-large, ResNet50, ResNet101)	Efficient Hybrid Encoder 사용	IoU-aware Query Selector Decoder	RT-DETR-X : 54.8% AP on COCO

II. TANGO 차별성

Detection 자동생성/학습 (Yolo History)



- 발체: YOLOv10: Real-Time End-to-End Object Detection, ArXive

신경망은 정확도와 Latency의 Tradeoff
신경망 알고리즘 개발 속도가 최적화에 의한 성능 향상 속도를 앞지르고 있음

II. TANGO 차별성

다양한 배포 환경 지원

학습은 생산성의 영역, 추론은 성능 최적화의 영역

- (다양한 타겟 환경 통합 지원) 구글, Nvidia, Intel 등 글로벌 기업들은 자사의 클라우드 혹은 자사 가속HW에 특화된 기술만 제공
- (실행 코드 자동 생성) 신경망 모델을 타겟 환경에서 실행하는데 필수적인 코드의 자동 생성 지원

◉ HW의 다양성 지원

- x86(windows, Linux), ARM 등 CPU 지원
- CUDA, NPU, ARM Mali, 퀄컴 Adreno 등 다양한 가속환경 지원

◉ 추론엔진의 다양성 지원

- PyTorch, NVIDIA TensorRT, 안드로이드 스마트폰 TensorFlow Lite 추론엔진 지원
- Apache 재단의 TVM, ARM사 ACL(Arm Compute Library) 추론엔진 지원

II. TANGO 차별성

다양한 배포 환경 지원

다양한 가속 환경, 다양한 추론엔진 지원

Target	NN acceleration	Runtime Engine	remark
Amazon Web Services (AWS)		PyTorch/TorchScript, TensorRT	Elastic Container Service (ECS)
Google Cloud Platform (GCP)		PyTorch/TorchScript, TensorRT	Google Cloud Run
KT cloud		PyTorch/TorchScript, TensorRT	in progress
Kubernetes	x86 + NVIDIA GPU	PyTorch/TorchScript, TensorRT	
PC	x86 + NVIDIA GPU	PyTorch/TorchScript, ONNX, OpenVINO, TensorRT, TVM	
Comma 3X (Snapdragon 845)	ARM + Adreno 630 GPU	PyTorch/TorchScript, ONNX	in progress
Jetson AGX Orin	ARM + NVIDIA GPU (Ampere)	TensorRT, PyTorch/TorchScript	
Jetson AGX Xavier	ARM + NVIDIA GPU (Volta)	TensorRT, PyTorch/TorchScript	
Jetson Nano	ARM + NVIDIA GPU (Maxwell)	TensorRT, PyTorch/TorchScript	
Samsung Galaxy S23	ARM + Adreno 740 GPU	Tensorflow Lite	
Samsung Galaxy S22	ARM + Adreno 730 GPU	Tensorflow Lite	
Raspberry Pi5	ARM + Google Coral M.2 PCIe TPU	Tensorflow Lite	
Odroid N2	ARM + Mali GPU	TVM, ACL	YoloV3 supporting
Odroid M1	ARM + RKNN NPU	RKNN	YoloV7 supporting

II. TANGO 차별성

다양한 배포 환경 지원

신경망 배포 모듈의 동작 예 (최적화 과정)

1) 스마트폰

PyTorch용 모델 -> ONNX모델로 변환 -> 정수형 양자화 -> Tensorflow Lite모델로 변환 -> 안드로이드용 신경망 실행 코드 생성
-> 안드로이드용 실행 파일(.apk) 생성

2) 라즈베리파이+TPU

PyTorch용 모델 -> ONNX 모델로 변환 -> Tensorflow Lite 모델로 변환 -> TPU용 모델로 변환 -> TPU용 Python 신경망 실행 코드 생성

3) TensorRT사용 디바이스

PyTorch용 모델 -> ONNX모델로 변환 -> ONNX-TensorRT변환기 및 TensorRT용 Python 실행 코드 생성

4) TVM사용 디바이스

PyTorch용 모델 -> ONNX 모델로 변환 -> TVM용 모델로 변환 -> TVM용 Python 신경망 실행 코드 생성

5) GCP/AWS 등 웹서비스

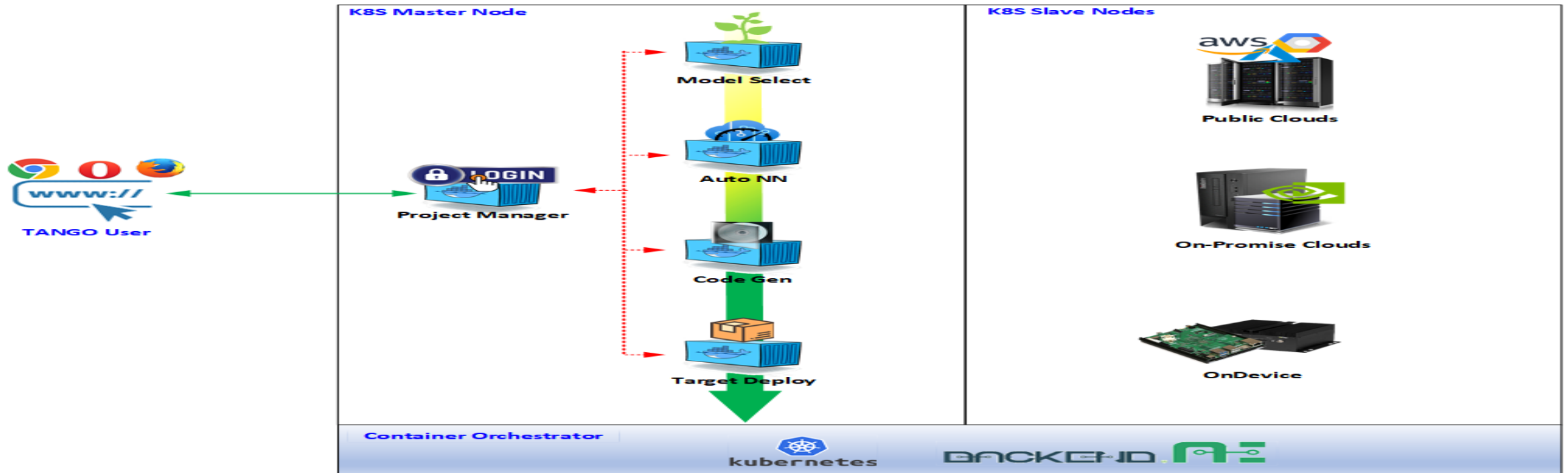
pytorch 모델 -> PyTorch기반 웹응용서비스 생성 -> K8S용 POD 생성 -> 사용자가 설정한 배포 환경으로 POD 배포/실행

II. TANGO 차별성

통합 파이프라인 지원

타겟 HW 인지형 신경망을 자동생성하고, 다양한 환경(클라우드·쿠버네티스 엣지·온디바이스)에 최적화된 배포까지 **하나의 도구로 자동화**

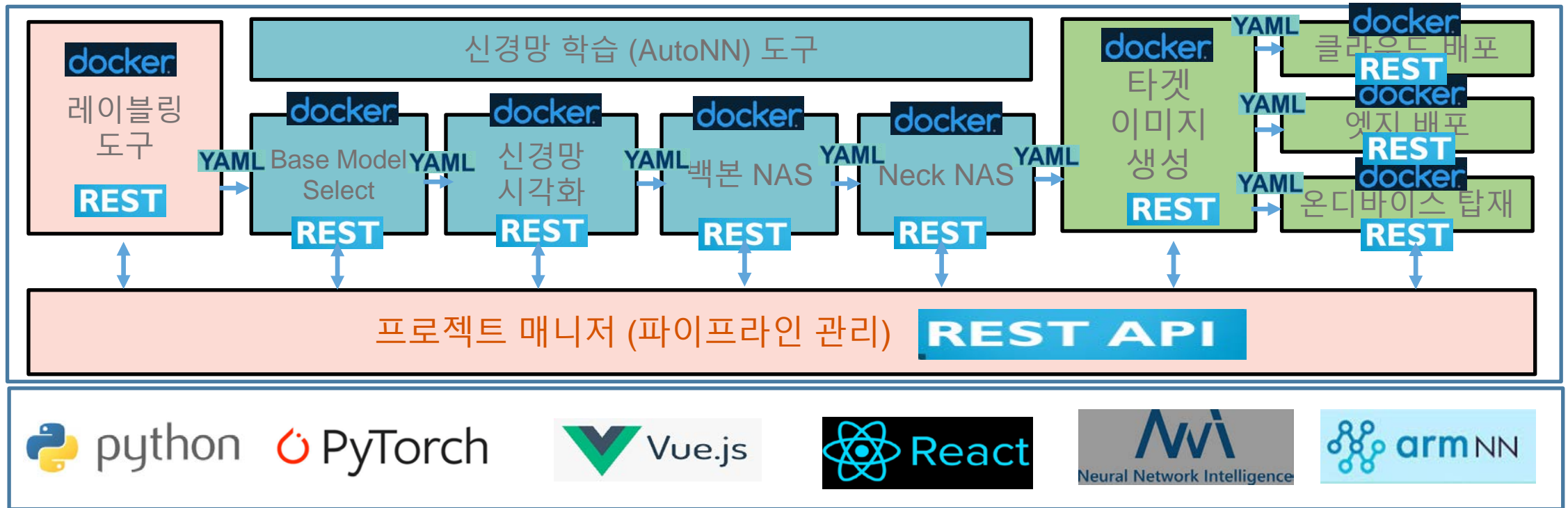
학습과 추론을 모두 지원
알고리즘 개발, 학습기법 연구, 배포탑재 추론엔진, AI컴퓨팅 시스템SW 전 분야에 대한 기술 개발



II. TANGO 차별성

Well Defined SW architecture

Docker 기반 MSA 구조, Rest API 통신, YAML 데이터 교환 정의



II. TANGO 차별성

Well Defined SW architecture

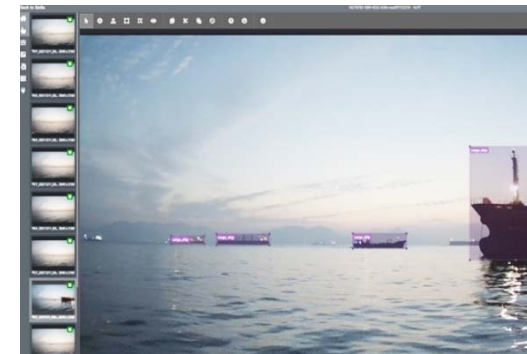
다양한 알고리즘 추가 및 향후 기능 확장에 최적화된 도커 구조

- **Tango as a Service (탱고 클라우드화)**
 - 쿠버네티스 연동을 고려
 - 국내외 클라우드사들(CSP) 대상으로 솔루션 사업화 고려
- **Web 기반 UI**
- **다양한 배포 환경을 고려**
 - 클라우드, 쿠버네티스 온프레미스, 온디바이스
 - 다양한 HW 가속(CPU, GPGPU(Cuda), NPU, ... 배포를 고려
- **모듈화 (서비스 컨테이너화)**
 - 시스템 분해 및 추상화하여 기능별로 분리
 - 성능향상, 시스템 수정, 재사용, 유지관리 편리
- **Multi-Node, Multi GPU 분산 학습**
 - Pytorch DDP를 통한 학습, 대용량 데이터를 위한 스토리지 관리, 노드간 통신 오버헤드 고려

Ⅲ. 성과 확산

실증을 통한 탱고 검증

- (주)웨다 : 스마트공장
 - 철강과 자동차 부품 제조 업체 2개 기업 대상 실증
 - 다양한 산업현장의 실제 데이터를 가지고 검증 및 사업화
 - TANGO 알고리즘을 Blue.ai에 내재화
- 서울대병원 : 의료
 - 흉부 X선 영상 데이터를 활용해 폐결핵 자동 검출 실증
 - 관상동맥 석회화 판별 인공지능 개발·검증
 - 골다공증 예측, 폐암 발생 위험 예측, 심혈관 질환 위험 예측 등 병의 예후 예측 기술 개발 계획 (DDPM 기반)
- (주)래블업 : 클라우드 사업화
 - 아마존 AWS, 구글 GCP 클라우드, 국산 KT클라우드 환경 자동 배포 지원
 - TANGO as a Service 검증
 - TANGO를 Backend.AI 기반으로 연동
- (주)에이브노틱스 : 스마트선박
 - 자율항해를 위한 온디바이스AI 사업화
 - 한국과학기술지주 투자 유치 (공공연구성과 확산 및 실용화사업)



Ⅲ. 성과 확산

기술이전 및 사업화

- 듀얼 라이선스를 통한 기술이전 (공개SW 과제의 사업화 모델 제시)
 - 연구용 사용 : GPL 라이선스, 코드 공개의무 있음
 - 사업용 사용: ETRI 기술이전, 코드 공개의무 없음
- ETRI Holdings 기술출자
 - 탕고 기술이전을 바탕으로 비즈니스 모델 확대 및 투자 병행
 - 기업가치의 10% 출자 조건
- 한국과학기술지주 기술출자
- 신용보증기금, 기술보증기금 대출
- NIPA 2024년 유망 SaaS 개발·육성 지원 사업 공고
 - 총 36.64억원 (총 8개 과제, 과제당 4.58억원 이내 지원)
 - 2024. 5 ~ 12 (8개월)

Ⅲ. 성과 확산

Future Work

- TANGO as a Service
 - TANGO를 클라우드 SaaS화
 - AI컴퓨팅에 최적화된 도커 컨테이너 컴퓨팅
- 병렬/분산 AI컴퓨팅 지원
 - 멀티노드, 멀티 GPU 학습 환경 구축
 - 대용량 도커 볼륨 처리, Pytorch DDP 처리, RAY 연동
- 지속적 학습 (CI/CD) 지원
 - 지속적 데이터 유입에 의한 학습 데이터 증가, 온디바이스의 가속 성능 향상
 - 신경망 모델의 진화 (동일 성능 대비 정확도 향상)
 - 효율적인 Continual Learning 알고리즘 개발
- 생성형AI 태스크 처리
 - 생성형AI에 MLOps화 및 다양한 산업 도메인에 적용
 - LoRA/DoRA, AI Agent, LLM/sLLM, RAG, Stable Diffusion

<https://github.com/ML-TANGO>

Tango는 공장, 의료 등 산업 분야에서
비교적 인공지능 전문지식이 부족한 사용자들에게
손쉬운 SW개발 프레임워크를 제공한다.

Tango는 개발 전 과정이 오픈소스로 공개되는 만큼,
많은 분들이 TANGO Github에 참여하여 주시기를 희망합니다.

감사합니다.

주관

ETRI
한국전자통신연구원

(TANGO)

주최



과학기술정보통신부

IITP

정보통신기획평가원

후원

labup

we-a

tesla
system

(사)한국인공지능협회
Korea Artificial Intelligence Association

SNUH
서울대학교병원

고려대학교
KOREA UNIVERSITY

홍익대학교
HONGIK UNIVERSITY

cau 중앙대학교
CHA UNIVERSITY