



# TANGO 커뮤니티 소개

성명 조창식

소속 ETRI

인공지능 기술의 대중화  
(AI Democratization)를 위한  
제2회 탱고 커뮤니티 컨퍼런스





1

## TANGO 프로젝트

1. Intro
2. 신경망 응용 개발의 어려움
3. 해외 MLOps 동향

2

## TANGO 차별성

1. Detection 자동 생성/학습
2. 다양한 배포 환경 지원
3. 타겟 디바이스 인지형 신경망 자동생성

3

## TANGO 공개SW

1. 모든 개발 과정을 Github에서
2. Well Defined SW architecture
3. 성과 확산
4. Future Work



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)

## TANGO 히스토리

### TANGO 히스토리

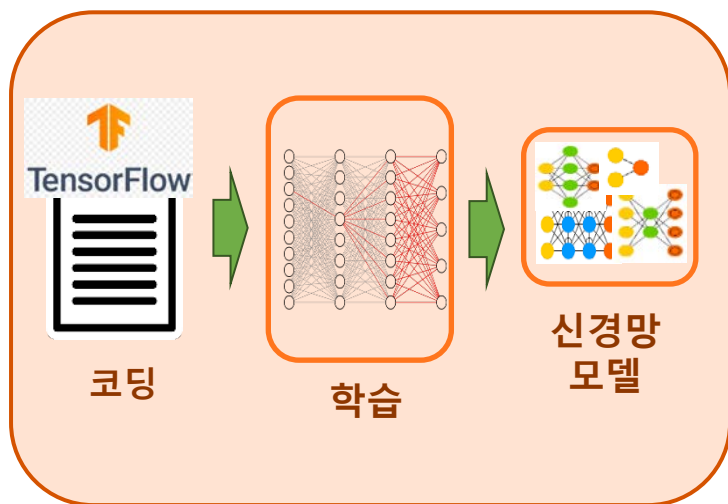
- 2021.04 신경망 자동생성 통합개발 프레임워크 과제 시작
- 2022.02 비공개 GitHub 저장소 운영
- 2022.09 공개 GitHub 저장소 운영
- 2022.10.31 Pre 릴리즈 (tango-22.11-pre1)
- **2022.11.01 1회 TANGO 커뮤니티 컨퍼런스 (AT센터 세계로룸)**  
94개 기관 158명이 참석
- 2022.11.30 22년 하반기 정식 릴리즈 (tango-22.11)
- 2023.05.31 23년 상반기 정식 릴리즈 (tango-23.06)
- 2023.10.23 23년 하반기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



## 신경망 응용 개발의 어려움



도메인 전문가



ML scientist



ML engineer



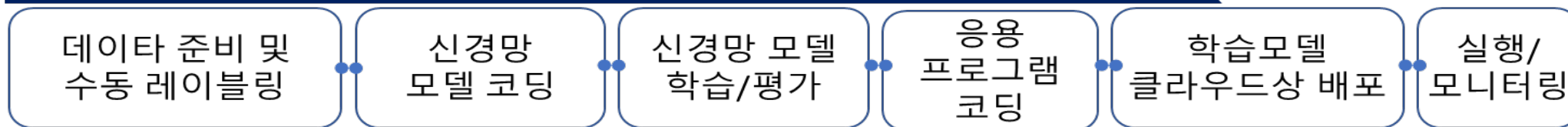
도메인 전문가

도메인 지식과 신경망 전문지식에 대한 고도의 개발경험 요구

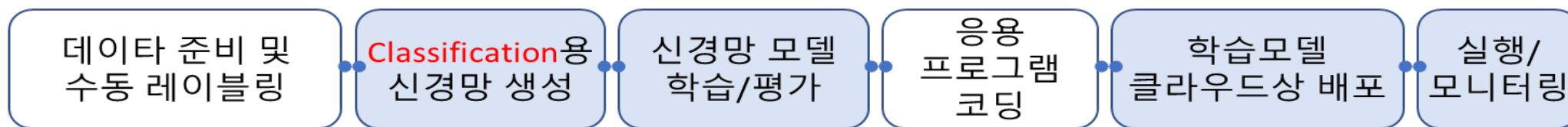


## 신경망 응용 개발의 어려움

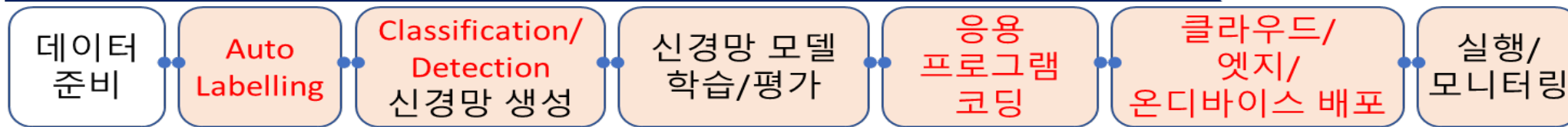
### 클라우드 기반 신경망 개발: 수동 코딩



### 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 Azur ML



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

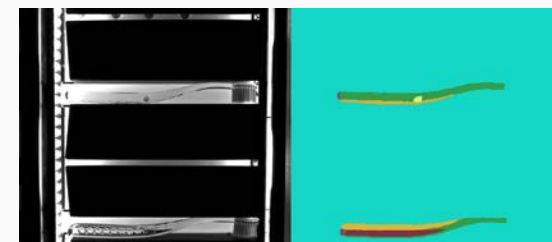
### Detection (용접불량 검사)



#### ○ 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)



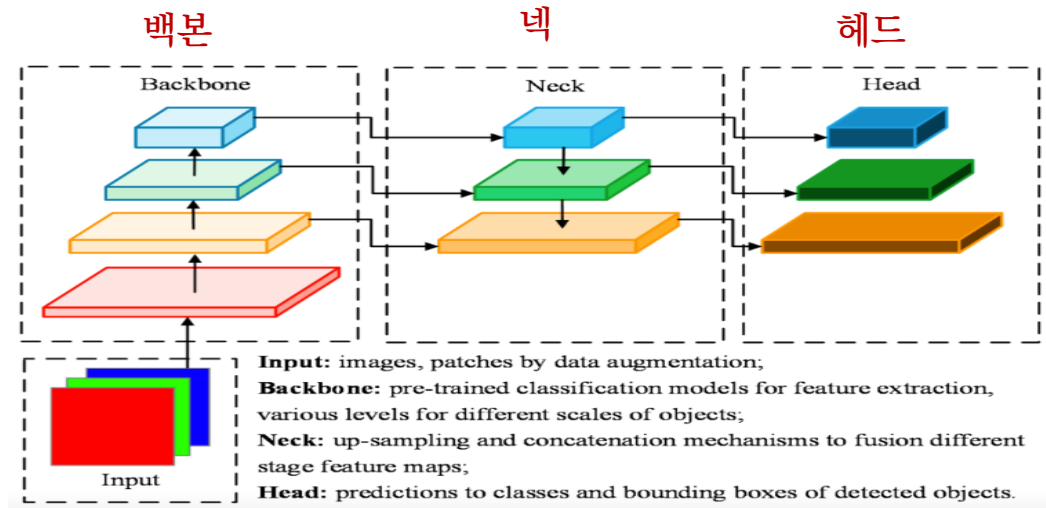
#### ○ 치솔 불량부위 검출

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



## Detection을 지원하는 신경망 자동생성 도구

Detection을 지원하는 신경망 자동생성 도구(세계최고 성능 추구)



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

넥 계층의 연결선에 대한 탐색을 통하여 정확도, 성능을 고려한 신경망 자동생성

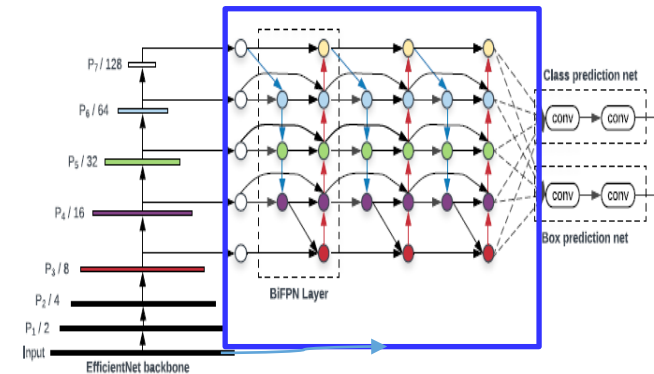


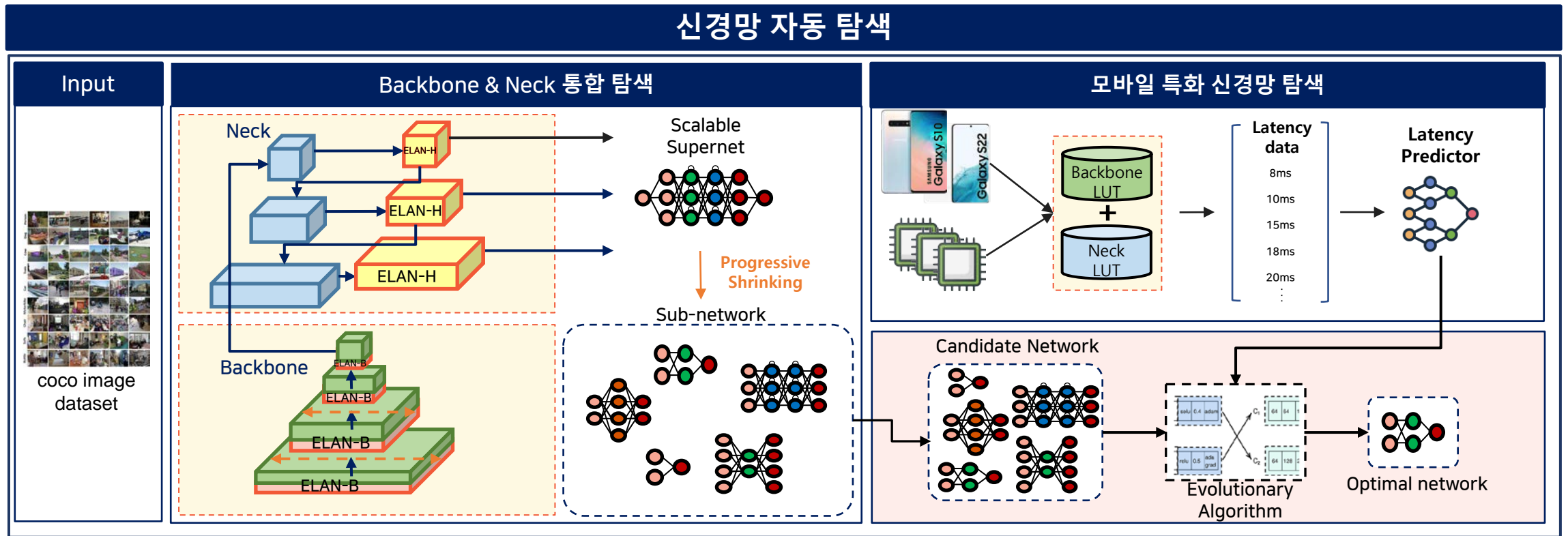
Figure 3: EfficientDet architecture - It employs EfficientNet [39] as the backbone network, BiFPN as the feature network, and shared class/box prediction network. Both BiFPN layers and class/box net layers are repeated multiple times based on different resource constraints as shown in Table 1.



### Detection을 지원하는 신경망 자동생성 도구

#### 백본과 넥 신경망 탐색을 통합한 신경망 자동 생성 기술

- 백본과 넥의 동시 탐색을 통해 객체 탐지 신경망의 구성 요소와 연결 구조를 고려하여 최적의 신경망 구조 탐색 가능





### 다양한 배포 환경 지원

다양한 배포환경, 다양한 추론엔진 지원

- (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

## 다양한 배포 환경 지원

다양한 가속 환경, 다양한 추론엔진 지원

Target Environment	Device Environment		Runtime engine
Cloud Environment 	GCP(Google Cloud Platform)		PyTorch
Kubernetes Environment 	x86 + NVIDIA GPU	PC	PyTorch
	ARM CPU + NVIDIA GPU	Jetson Nano	TensorRT
	ARM CPU + NVIDIA GPU 	Jetson AGX Orin	TensorRT, PyTorch
		Jetson AGX Xavier	TensorRT, PyTorch
		Jetson Nano	TensorRT, PyTorch
	ARM CPU + Adreno GPU 	S22	Tensorflow Lite(Android)
OnDevice Environment 	ARM CPU + Mali GPU	Odroid-N2	TVM, ACL



## 타겟 디바이스 인지형 신경망 자동생성

- 타겟 적응형 신경망 자동 생성을 위한 세밀한 신경망 모델 추천
  - (현재) 사용자가 직접 디바이스의 정보 입력 및 등급 설정
  - (추가) NLP 이용한 디바이스의 사양 자동 설정 기능
  - (추가) 기존에 학습했던 결과에 따른 추천 기능 반영

CPU/제작업체

Processor	CPU Cores	AI Accelerator	Year	CPU Q1 Score	CPU F1 Score	Q1 Accuracy	Q1 Accuracy	FP16 Score	FP16 Accuracy	FP32 Score	FP32 Accuracy	AI Score
Unisc Tiger T120	4x2.4GHz Cortex-A75 & 4x1.8GHz Cortex-A55	NPU / na.	2019	1400	2113	8455	96	14603	79	391	384	86
Snapdragon 855 Plus	1x2.96 + 3x2.42 + 4x1.8GHz Kryo 485	DSP (Hex 650) - GPU (Adreno 640)	2019	2213	4132	6375	40	8960	30	1357	1078	34
HiSilicon Kirin 810	2x2.27GHz Cortex-A76 & 6x1.88GHz A55	NPU / na.	2019	1085	2488	1895	80	17635	95	403	348	90
Snapdragon 855	1x2.84 + 3x2.41 + 4x1.78GHz Kryo 485	DSP (Hex 650) - GPU (Adreno 640)	2018	2106	3431	5535	55	7342	37	1156	714	43
Mediatek Helio P90	2x2.2GHz Cortex-A75 & 6x2GHz Cortex-A55	DSP x2 + APU / na.	2019	1067	2018	5666	98	10120	94	140	69	96
Exynos 9825 Octa	na	GPU / na + NPU x2	2019	1491	2396	2077	60	9170	46	780	717	50
HiSilicon Kirin 980	2x2.6GHz + 2x1.52GHz A76 & 4x1.8GHz A55	NPU x2 / na.	2018	1817	3447	222	60	10750	85	139	64	76
Exynos 9820 Octa	2x2.7GHz M4 & 2x2.3GHz A75 & 4x2GHz A55	GPU (Mail-G76 MP21) + NPU x2	2018	1491	2545	1950	60	8367	45	744	701	50
Snapdragon 845	4x2.8GHz Kryo 385/G & 4x1.7GHz Kryo 385/S	DSP (Hex 685) - GPU (Adreno 630)	2018	1605	2172	2031	58	6327	38	916	683	45
Snapdragon 730	2x2.2GHz Kryo 470/G & 6x1.8GHz Kryo 470/S	DSP (Hex 688) - GPU (Adreno 618)	2019	1060	2082	2765	58	3581	37	450	374	44

AI가속 GPU

성능

타겟 디바이스

순번	Device	응용	Data Set	Resolution	Scratch /Transfer	Neural Network	Accuracy	FPS
1	D3	객체인지	CoCo	640*640	S	Yolov7x	53	24
2	D0	객체인지	Racing	480*400	T	Yolov7-tiny	39	10
3	D1	객체인지	Racing	480*400	S	Yolov7-tiny	39	15
4	D4	객체인지	CoCo	640*640	S	Yolov7-E6E	56.7	100
5	D1	객체인지	볼트	320*320	T	Yolov7-tiny	52	30
6	D0	분류	CIFAR10	320*320	S	Resnet-18	85	13
7	D1	분류	Inspection	160*160	T	Resnet-24	90	20
8	D2	분류	ImageNet	320*320	S	Resnet-50	68	40
9	D							
⋮	⋮	⋮	⋮					
⋮	⋮	⋮	⋮					
10000	D2							
10001	D3							

상세 정보는 NLP 이용

최적 신경망

추론 수행 후 리턴값





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>

모든 개발 과정을 Github에서

매년 두 번의 릴리즈 버전을 완성하고 하반기에 공개SW 세미나 추진





## Well Defined SW architecture

### Docker 기반 MSA 구조

#### 요구사항 분석

- ◉ Web 기반 UI
- ◉ 다양한 기술 스택 지원  
언어: Python, C++, ...  
딥러닝 프레임워크: PyTorch, Tensorflow
- ◉ 다양한 HW 타겟  
온디바이스, 클라우드(온프레미스)  
CPU, GPGPU(Cuda), NPU, ...
- ◉ 버전 관리, 이슈관리

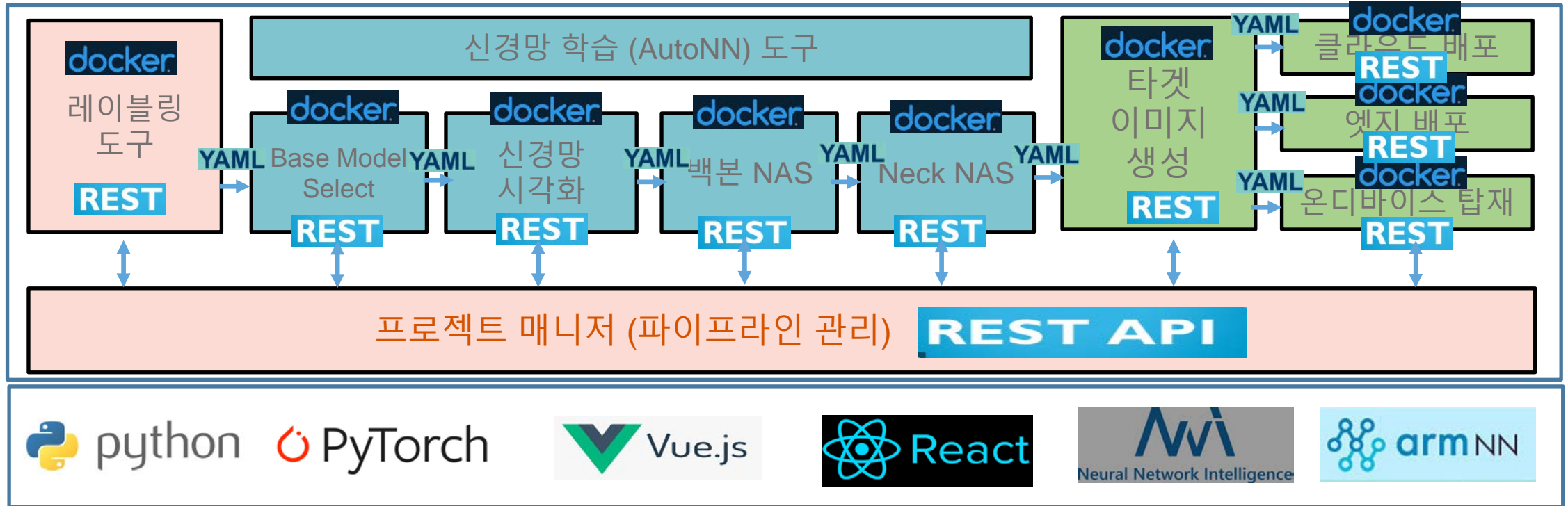
- ◉ 모듈화(서비스 컨테이너화)
  - 시스템 분해 및 추상화하여 기능별로 분리
  - 성능향상, 시스템 수정, 재사용, 유지관리 편리

- ◉ MSA Benefits (출처: <https://microservices.io>)
  - Highly maintainable and testable
  - Loosely coupled
  - Independently deployable
  - Organized around business capabilities
  - Owned by a small team



#### Well Defined SW architecture

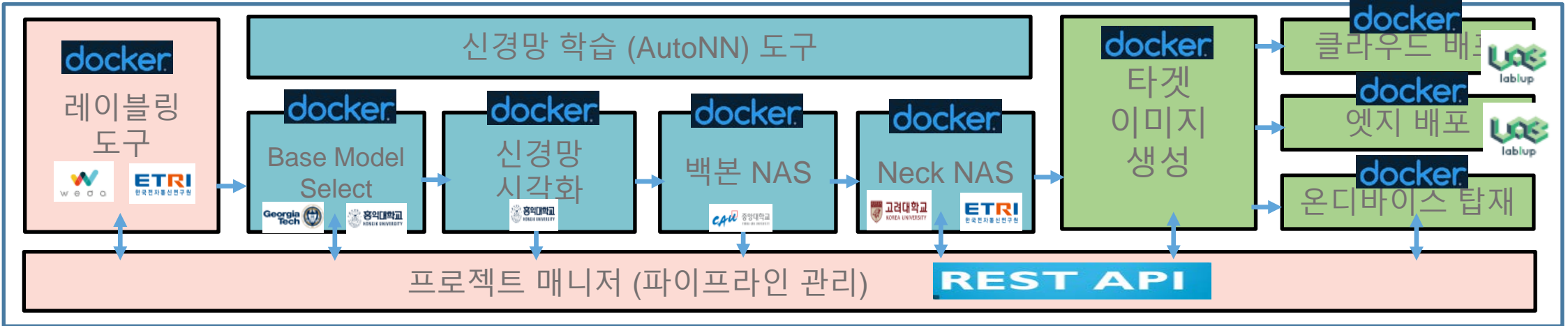
Docker 기반 MSA 구조, Rest API 통신, YAML 데이터 교환 정의





## Well Defined SW architecture

다양한 알고리즘 추가 및 향후 기능 확장에 최적화된 MSA(도커) 구조



### 알고리즘 다양화

- ◉ **Base Model Select (다중 Approach 접근)**
  - ETRI (태스크 기반 Rule-based 제안)
  - 조지아공대 (Feature Engineering 기반 제안)
- ◉ **Neck NAS (다중 알고리즘 접근)**
  - 중앙대(All-in-one), 고려대(Porxyness NAS), ETRI (SuperNet NAS) 병렬

### 기능 확장

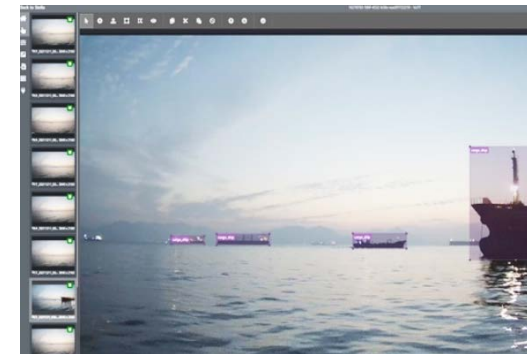
- ◉ 드래그 & 드랍 파이프라인 관리
- ◉ Multi-Node Multi GPU 분산 학습
- ◉ Backend.ai(래블업) 오케스트레이션
- ◉ Tango as a Service (탱고 클라우드화)
- ◉ 국내 CSP 솔루션화



#### 성과 확산 (실증 및 산업적용)

- (주)웨다 : 스마트공장
  - TANGO 알고리즘을 Blue.ai에 내재화
  - 다양한 산업현장의 실제 데이터를 가지고 검증 및 사업화
- (주)래블업 : 인프라
  - TANGO as a Service (TaaS) PoC
  - 클러스터 세션 기반의 멀티노드 연산 지원
  - TANGO를 Backend.AI 기반으로 연동
- (주)에이브노틱스 : 스마트선박
  - 스마트선박에 TANGO AutoML 알고리즘 및 프레임워크 사용
  - 스마트선박용 온디바이스 배포에 적용
- 서울대병원 : 의료
  - 서울대병원 실제 데이터에 TANGO 알고리즘 적용
  - 병의 예후 예측 기술 개발 (DDPM 기반)

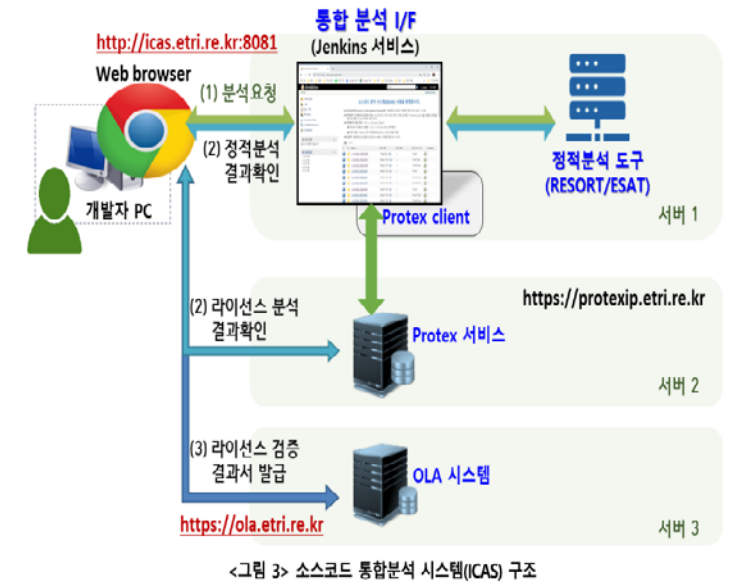
BLUE AI





#### 성과 확산 (라이선스 및 품질관리)

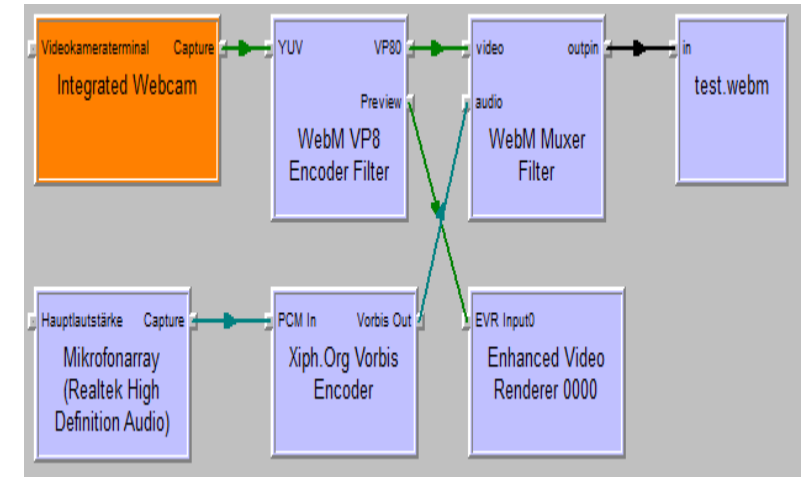
- 듀얼 라이선스를 통한 기술이전 (공개SW 과제의 사업화 모델 제시)
  - 연구용 사용 : GPL 라이선스, 코드 공개의무 있음
  - 사업용 사용: ETRI 기술이전, 코드 공개의무 없음
- 오픈소스 라이선스 검증 및 코드 품질 관리
  - TANGO 코드 릴리즈마다 검증
  - 탱고 라이선스와 사용하는 공개 라이브러리의 라이선스 충돌 해결
  - ETRI ICAS (Integrated Code Analysis System), Protex 사용하여 코드 품질 관리 및 검증
- 인공지능 개발시 발생하는 라이선스 이슈 해석 중
  - 코드 사용, 알고리즘 사용, Model Zoo, Transfer Learning





## Future Work

- TANGO as a Service
  - TANGO를 구글 GCP 클라우드 GKE 엔진으로 SaaS화
  - TANGO 프로젝트내에 Repository 개설, Pilot으로 진행 (래블업)
- 파이프라인 편집 도구
  - TANGO의 다양한 AutoML 알고리즘을 그래프 편집으로 동적 파이프라인화
  - CI/CD 지원
  - Ex) MS directshow 편집 도구
- 분산 컴퓨팅 지원
  - 멀티노드, 멀티 GPU 학습 환경 구축
  - 연합학습 기능 TANGO 결합
- 생성형AI 태스크 처리
  - 생성형AI에 대한 AutoML 적용 및 MLOps화
  - 다양한 산업 도메인에 적용



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

Tango는 AI·SW 지식이 부족한 타 산업에서도  
AI 기반 SW를 손쉽게 개발할 수 있도록 함으로써  
AI·SW 기술의 전 산업 확산·디지털 혁신이 촉진될 것이다.

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



감사합니다.

