




# 신경망 배포 탑재 기술

성명 이경희

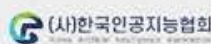
소속 ETRI

## SUBJECT

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

주관 ETRI (  ) 주최  과학기술정보통신부  정보통신기획평가원

후원





## 목 차

### 1 신경망 배포 탑재 기술의 개요

3

### 2 기술 개발 현황

6

### 3 향후 계획

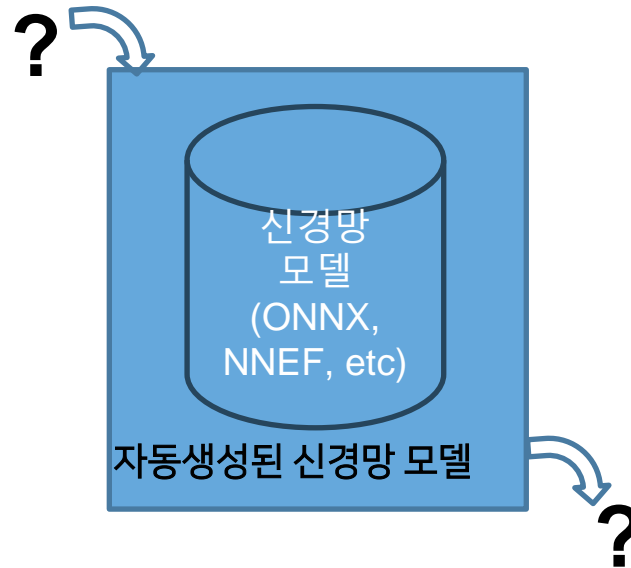
18

## 필요성

- 신경망 모델만 있는데 어떻게 실행시키는가?
- 신경망 모델에 입력은 어떻게 주어야 하나?
- 신경망 모델의 출력은 무엇인가? 어떻게 사용?



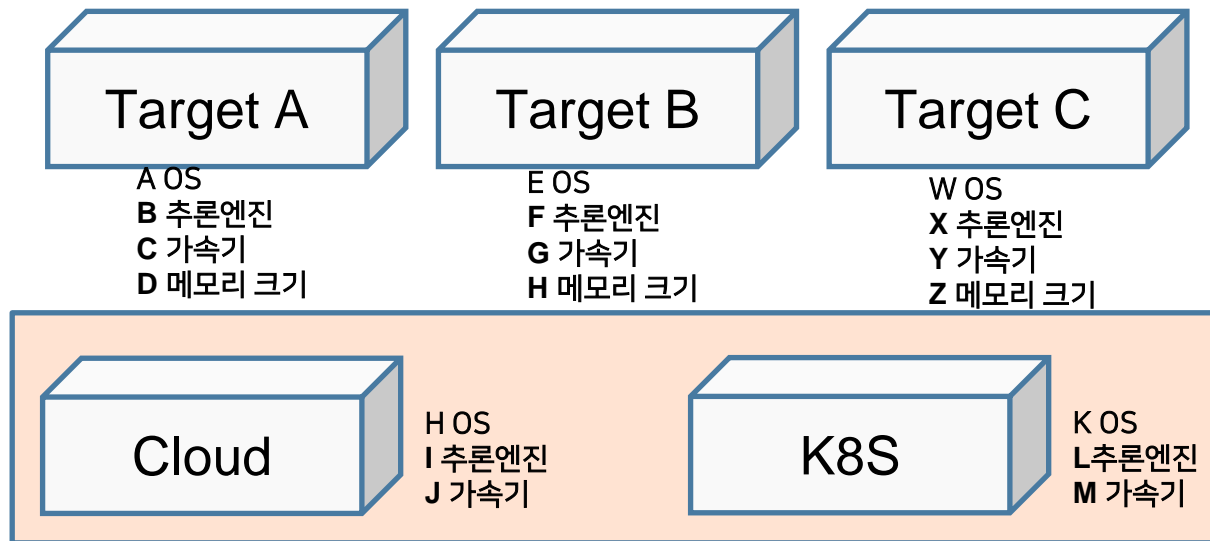
- **타겟 디바이스 맞춤형 신경망 응용 생성 지원**
  - 추론엔진 의존적 실행 코드 자동 생성
  - 신경망별 맞춤형 신경망 입출력 코드 자동 생성



# I. 신경망 배포 탑재 기술의 개요

## 필요성

- 온디바이스들은 실행 환경이 서로 다른데 어떻게 포팅해 실행하는가?
- 학습 신경망 엔진과 타겟의 추론 엔진이 다르면 어떻게 사용?
- 클라우드, 웹서버 등에서 실행시키려면 코딩은 어떻게 해야 하나?
- K8S 등을 통해 배포하는데 절차 및 방법은 어떻게?



## ○ 다양한 추론엔진 지원

- PyTorch, NVIDIA TensorRT
- TensorFlow, TensorFlow Lite
- ARM ACL(Arm Compute Library), Apache 재단의 TVM

## ○ 신경망 가속기의 다양성

- x86, ARM 등의 CPU
- NVIDIA GPU, ARM Mali 등 GPU
- 기타 NPU 등

## ○ 배포 및 탑재의 편의성 제공

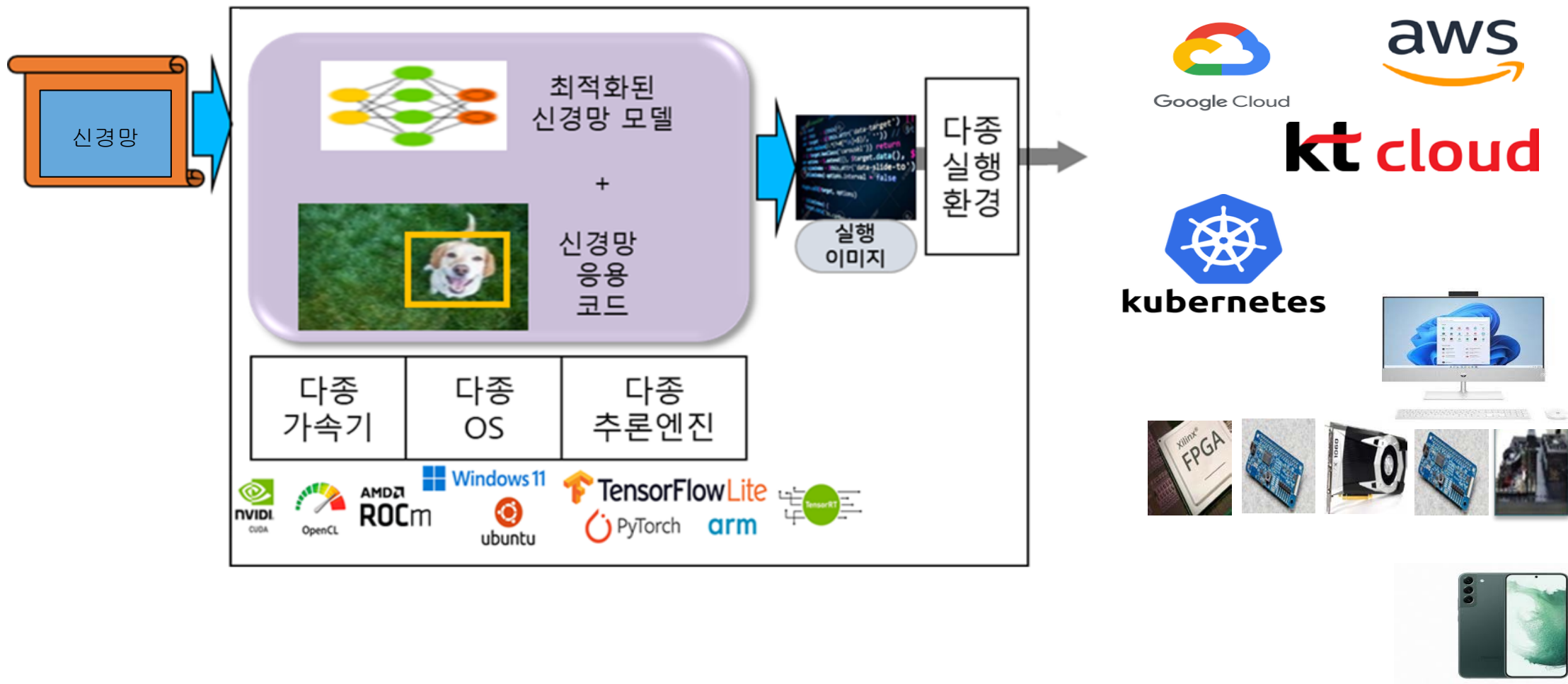
- cloud, K8S상 배포 지원 기능
- docker, 실행 파일 자동 생성 기능

타겟의 연산 환경/추론엔진/실행 환경별 신경망 모델의 최적화, 추론용 실행코드 생성, 타겟으로 배포 및 실행의 자동화 필요

# I. 신경망 배포 탑재 기술의 개요

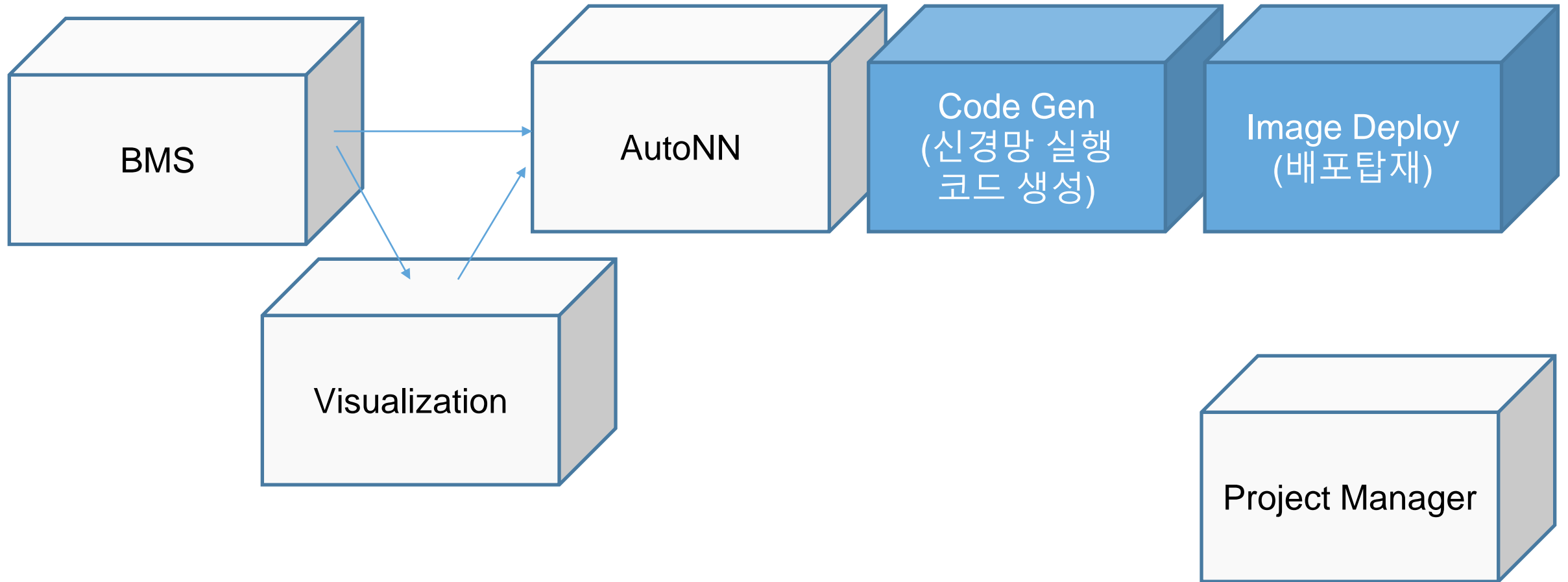
5

## 동작 구조





### 통합개발 프레임워크



### 기능 요약

#### ○ 신경망 생성 모델의 실행을 위한 전처리/후처리 코드 생성

- 전처리: 이미지 resize/crop, 평균치 조정, 신경망 입력용 텐서 생성(NCHW)
- 후처리: 신경망 모델의 출력 해석 (인식 객체명, 확률값)

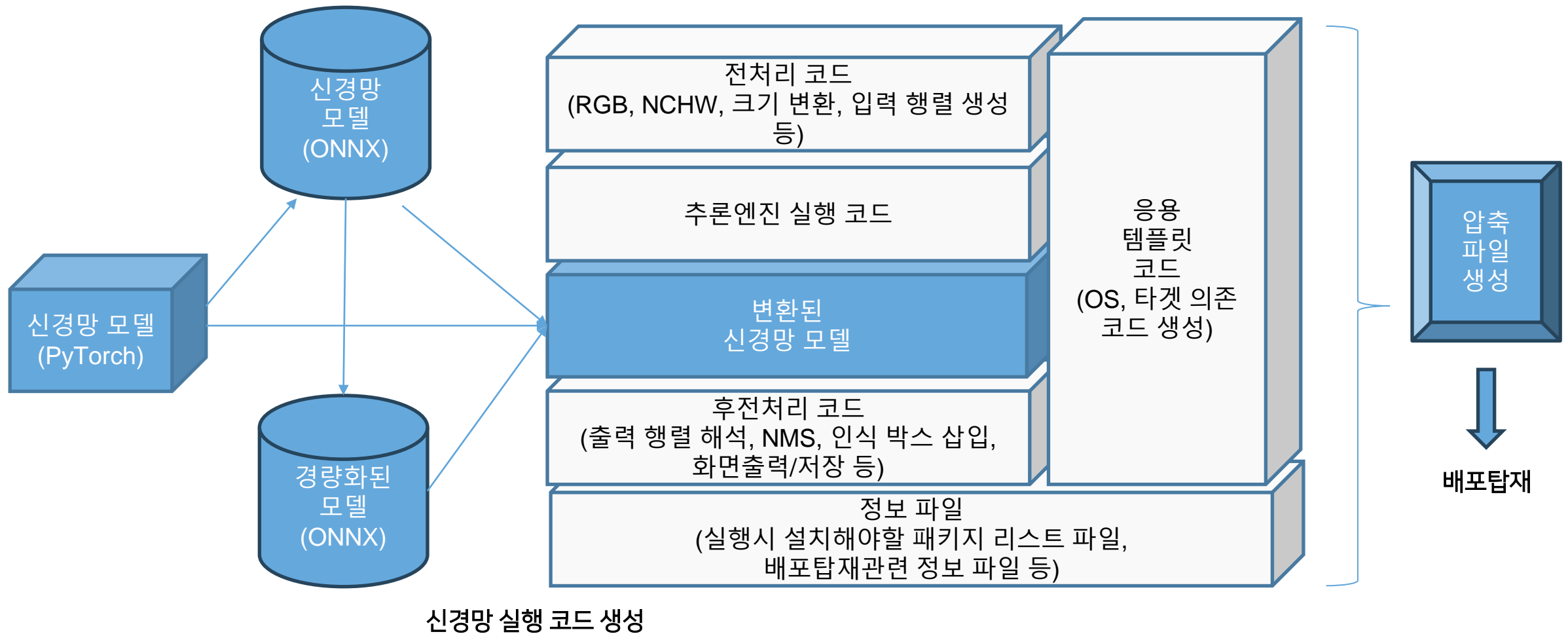
#### ○ 신경망 실행을 위한 입출력 코드 생성

- 입력: 파일/동영상/카메라 입력 코드 생성
- 출력: 동영상, 화면, 텍스트 등으로 결과 출력 코드 생성

#### ○ 배포탑재 및 실행을 위한 정보 파일 해석 및 생성

- AutoNN과 Project manager에서 생성한 yaml파일 해석
- 클라우드/엣지클라우드/온디바이스상 신경망 배포 실행을 위한 yaml 파일 생성

### 동작 흐름도

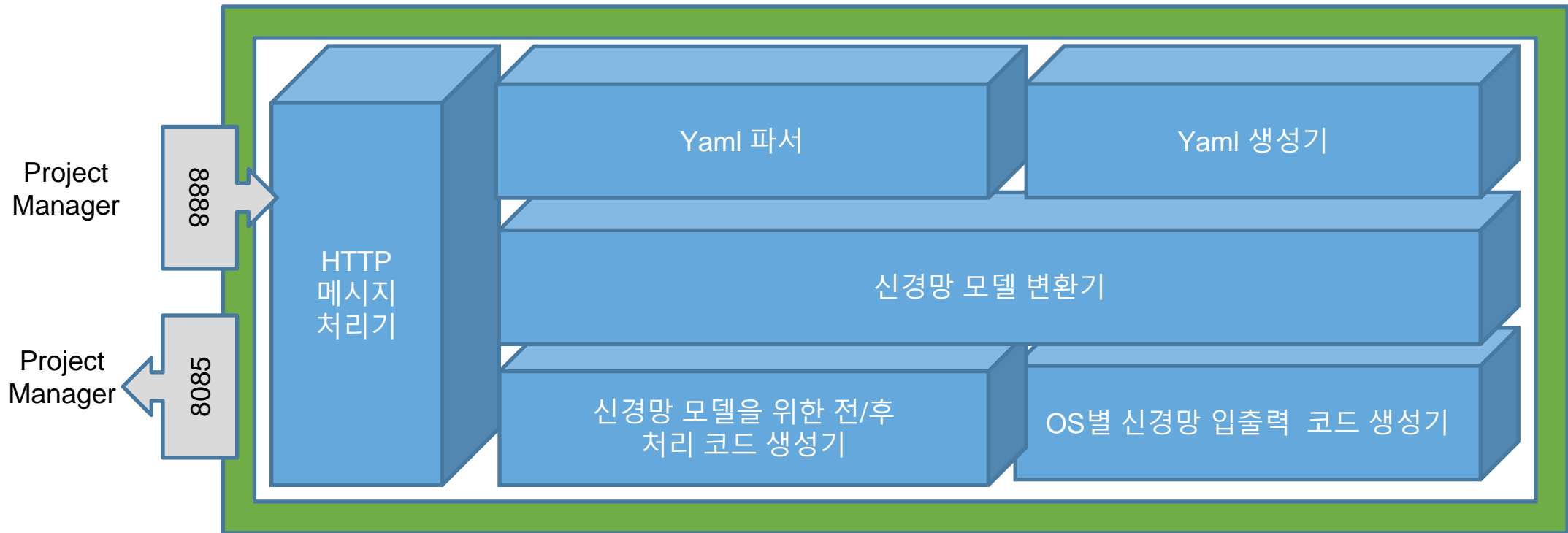




### 소스 코드 구조

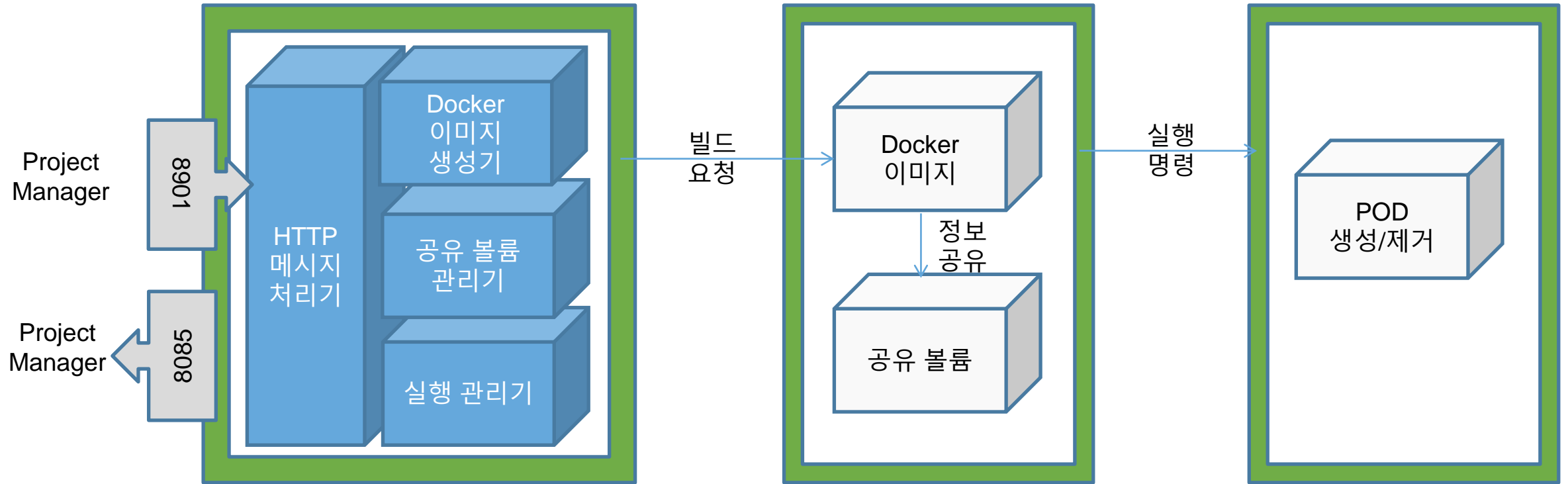
폴더명			모듈명
/TANGO	deploy_codegen	optimize_codegen	신경망 실행 코드 생성 모듈
	deploy_targets	cloud	클라우드 배포 모듈
		k8s	K8s 배포 모듈
		ondevice	온디바이스 배포 모듈

### 블록 구성도(optimize\_codegen)





### 블록 구성도(K8s)



### 블록 구성도(ondevice)





### 지원 현황

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

### 동작예: Kubernetes

#### ○ 신경망 실행 코드 생성 모듈

- PyTorch 기반 Web용 신경망 실행 코드생성
- 타겟 디바이스의 IP 정보 등



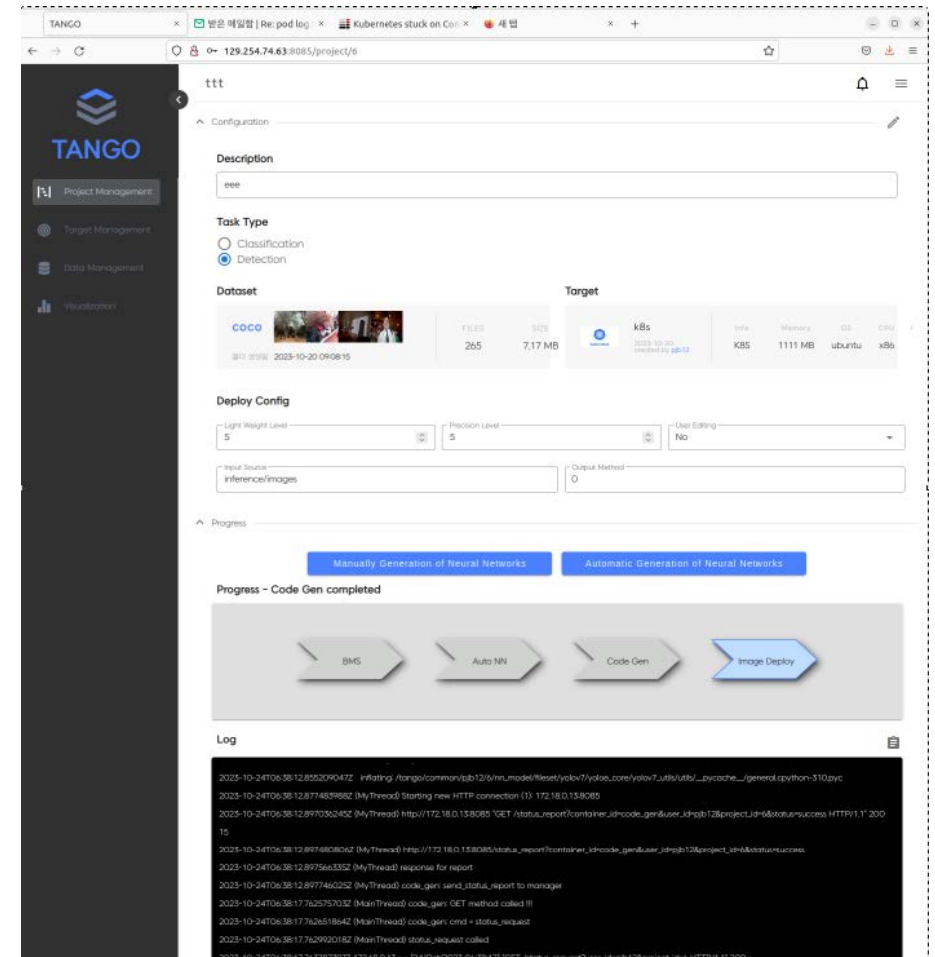
#### ○ K8S 배포 모듈

- 원격 타겟디바이스에 쿠버네티스 환경 구축
- 원격 타겟디바이스에 신경망 구동 환경 구축
- 원격에서 신경망 실행 명령 전달



#### ○ 타겟 시스템

- 신경망 실행
- 원격시스템에서 신경망 구동 결과 확인 가능





### 동작예: 갤럭시 S22

#### ○ 신경망 실행 코드 생성 모듈

- PyTorch -> ONNX 변환
- ONNX 모델 양자화
- TensorFlow 모델로 변환
- TensorFlow Lite 모델로 변환
- 안드로이드용 코드 생성  
(이미지 입력, 전처리 코드, 후처리 코드, NMS 코드, 출력 코드 등)
- 안드로이드용 응용 생성

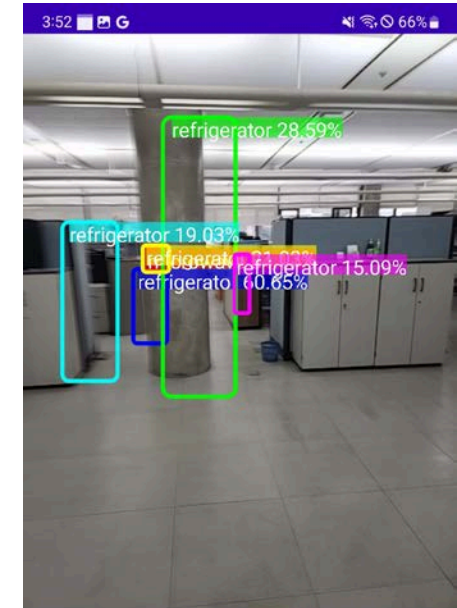


#### ○ 온디바이스 배포 모듈

- 안드로이드 응용 프로그램 다운로드



#### ○ 갤럭시 S22 스마트폰: 응용 프로그램 설치 후 실행



BaseModel: yoloe, RunMode: NONE\_FP32, InputSize: 640  
FPS: 6.17  
inference: 87ms postprocess: 3ms  
Confidence Threshold: 0.25  
IoU Threshold: 0.45



### 동작예: NVIDIA AGX Orin 보드

#### ○ 신경망 실행 코드 생성 모듈

- PyTorch -> ONNX 변환
- TensorRT용 실행 코드 생성  
(ONNX2TensorRT 모델 변환 코드,  
이미지 입력, 전처리 코드, 후처리 코드, NMS 코드, 출력 코드 등)



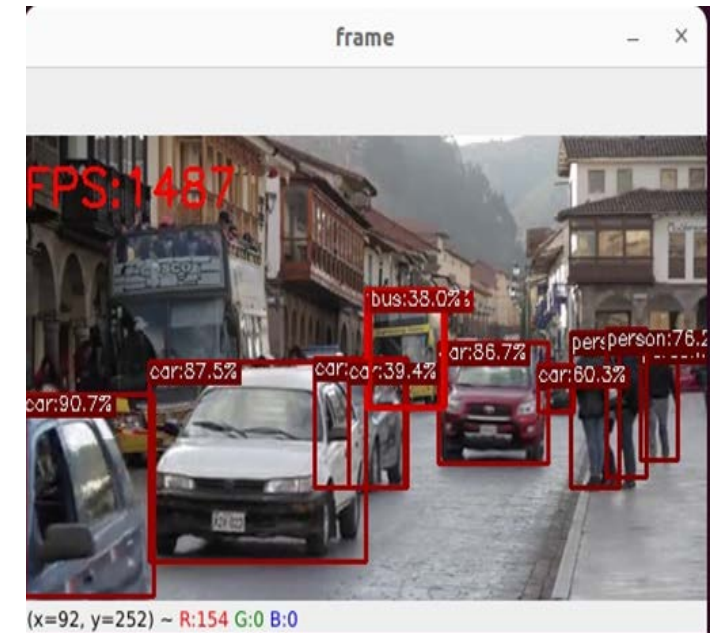
#### ○ 온디바이스 배포 모듈

- TensorRT용 응용 코드 다운로드



#### ○ NVIDIA AGX Orin

- 실행시 필요한 패키지 설치
- TensorRT용 응용 실행  
(ONNX2TensorRT 모델 변환, 객체 인식 기능 실행)





### 동작예: PC(Classification)

#### ○ 신경망 실행 코드 생성 모듈

- PyTorch 모델 입력
- PyTorch용 실행 코드 생성  
(이미지 입력, 전처리 코드, 후처리 코드, 출력 코드 등)



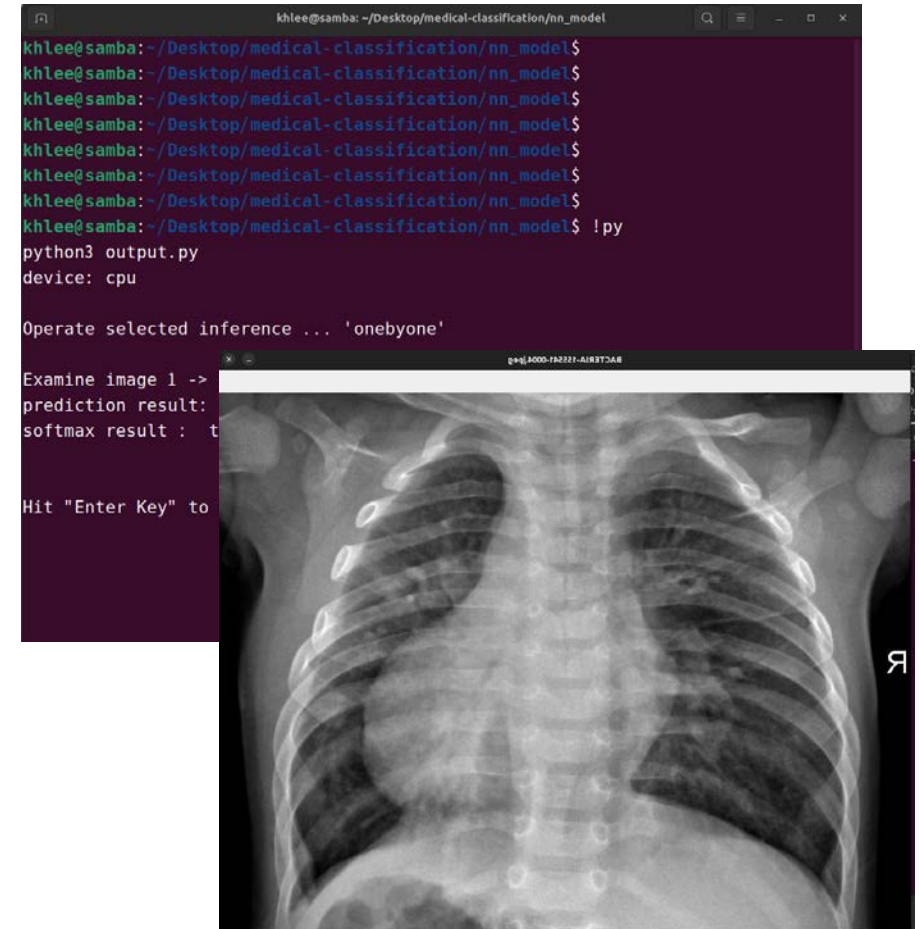
#### ○ 온디바이스 배포 모듈

- PyTorch용 응용 코드 다운로드



#### ○ PC

- 실행시 필요한 패키지 설치
- PyTorch 응용 실행



## Ⅲ. 향후 계획

### Future Work

- 지원 타겟 환경 확대
  - SDV
  - XR
- 고속 실행 기술 지원 강화
  - 국산 AI 가속기 지원
  - 신경망 고속 분산 실행 기술 지원 등
- 최신 신경망 기술 지원
  - 언어모델 등 멀티모달 지원
  - Diffusion 모델 등 생성형 AI 지원
- 적용 산업 영역 확대



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