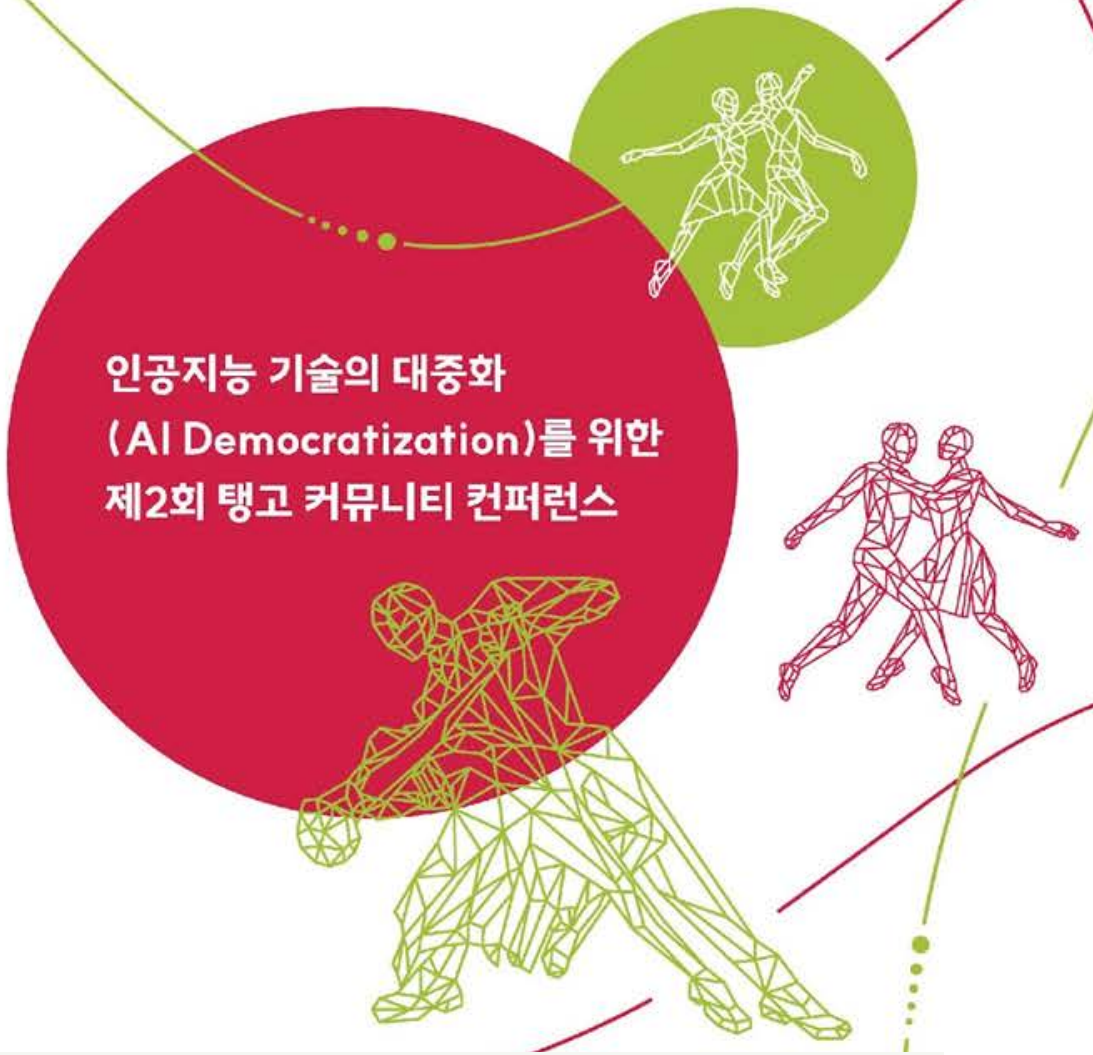
TANGO

신경망 자동생성 기술



성 명 김선태 소 속 한국전자통신연구원

주관 ETRI 한국전자통신연구원 (TANGO) 주최 과학기술정보통신부 IITP 정보통신기획평가원
후원 Lablup weda tesla System (사)한국인공지능협회 SNUH 서울대학교병원 고려대학교 KOREA UNIVERSITY 홍익대학교 HONGIK UNIVERSITY 중앙대학교 Georgia Tech

목 차

Ⅰ. 신경망 통합개발 프레임워크(TANGO)

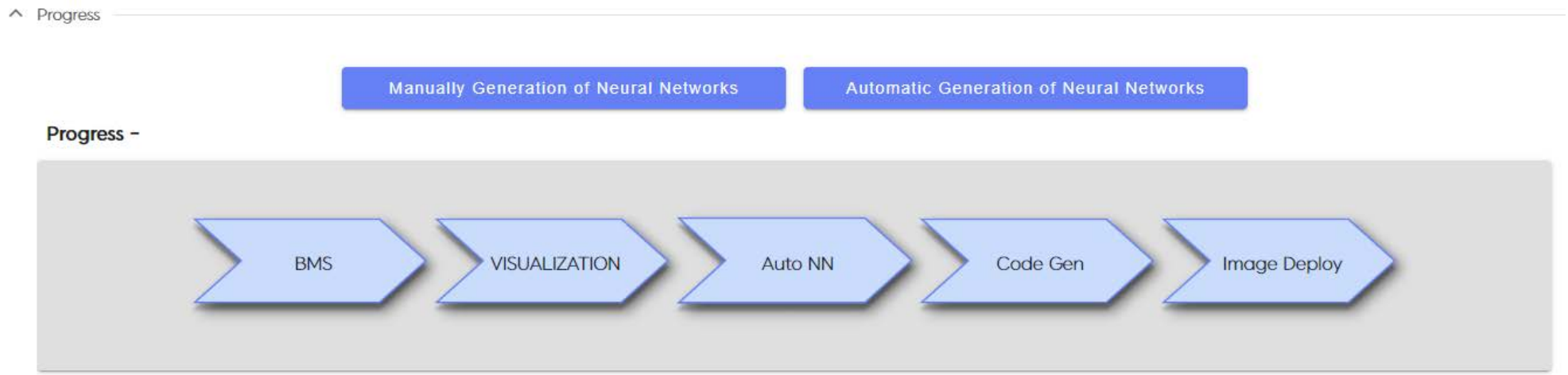
① 타겟 적응형 신경망 자동 생성도구 ② 타겟 맞춤형 신경망 응용 최적배포 도구

1. 신경망 통합개발 프레임워크(TANGO) - 모듈구성도

① 타겟 적응형 신경망 자동 생성도구 ② 타겟 맞춤형 신경망 응용 최적배포 도구

- **타겟 적응형 신경망 자동 생성**
  - (정의) 개발자 요구 사항에 따라 최적 신경망 생성
  - (요소) 신경망 모델 추천, 신경망 시각화, 신경망 생성(AutoNN)

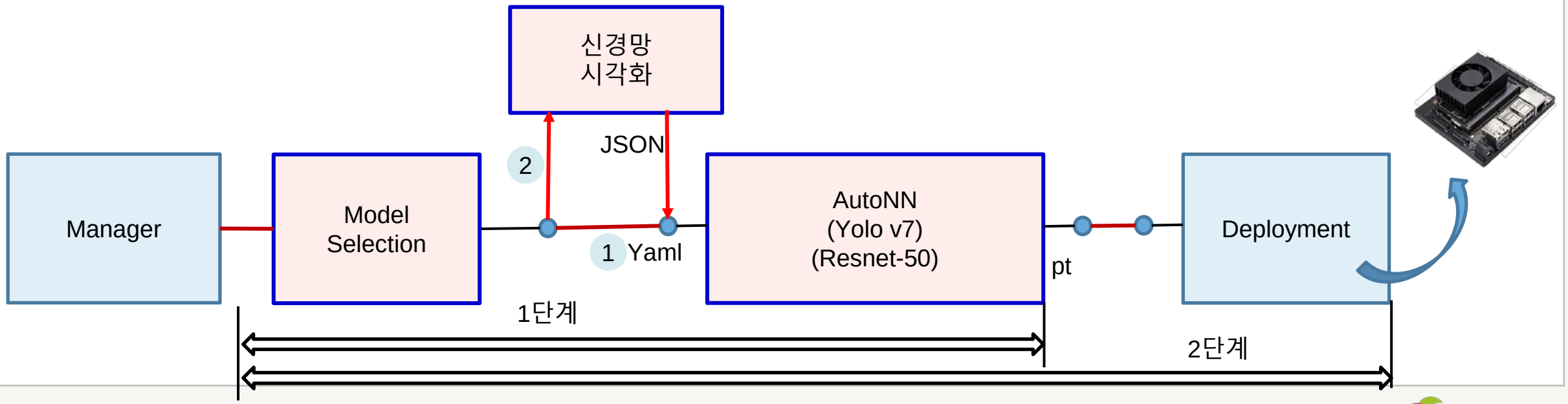
- **타겟 맞춤형 신경망 응용 최적배포 도구**
  - (정의) 추론 디바이스 환경과 성능에 최적화된 신경망 생성 및 실행
  - (요소) 신경망 생성 및 신경망 배포



2. 신경망 자동 생성 – 구성 모듈

① 타겟 적응형 신경망 자동 생성도구 ② 타겟 맞춤형 신경망 응용 최적배포 도구

- **타겟 적응형 신경망 자동 생성**
  - **신경망 모델추천 : 디바이스 사양과 응용에 적합한 모델 추천하는 과정**
  - **신경망 시각화 : 추천한 신경망에 대해서 사용자가 세부 수정하는 과정**
  - **신경망 생성 (AutoNN) : 데이터 세트를 이용하여 신경망 학습하는 과정**
    (이미지 분류 및 객체 탐지 신경망 제공)



2. 신경망 자동 생성 – 신경망 모델 추천

응용에 따른 신경망 종류와 디바이스 성능에 따른 복잡도로 신경망 추천

- **데이터 세트 설정**
  - **이미지 분류**(Image Classification) : **의료 데이터**, ImageNet, CIFAR10/100 **등**
  - **객체탐지** (Object Detection): COCO
- **타겟 디바이스 설정 (산업 적용 디바이스)**
  - **8개 디바이스 종류 지원**
  - CPU(ARM, Intel), GPU(nvidia, Mail **등**), NPU **등의 가속기 지원**

2. 신경망 자동 생성 – 신경망 모델 추천

응용에 따른 신경망 종류와 디바이스 성능에 따른 복잡도로 신경망 추천

- **이미지 분류 4종류 복잡도와 객체 탐지 6종류 복잡도로 추천**

Target		BMS output	
		Detection(6)	Classification(4)
Cloud	Cloud	Yolov7_E6E	Resnet152
K8S	K8S	Yolov7_W6	Resnet152
	K8S_Jetson_Nano	Yolov7_Tiny	Resnet34
PC	PC_Server	Yolov7_E6	Resnet152
	PC	Yolov7_W6	Resnet152
On Device	Jetson_AGX_Orin	Yolov7_W6	Resnet101
	Jetson_AGX_Xavier	Yolov7_X	Resnet50
	Jetson_Nano	Yolov7_Tiny	Resnet34
	Galaxy_S22	Yolov7_Tiny	Resnet34
	Odroid_N2	Yolov7_Tiny	Resnet34

추천된 신경망 모델에 대해서 복잡도 증감 및 파라미터 세부 튜닝 기능 제공

- **신경망 시각화 지원 모델**
  - VGG, Resnet **등 신경망** Backbone **시각화**
  - Yolo **모델 등 객체 탐지 지원**

신경망 수정

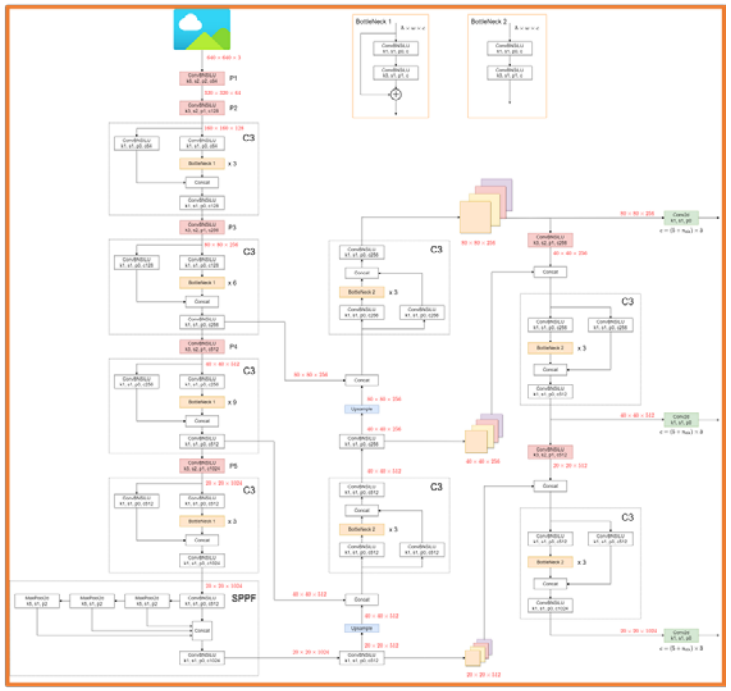
```
# YOLOv5 by Ultralytics, GPL-3.0 license

# Parameters
nc: 80 # number of classes
depth_multiple: 1.0 # model depth multiple
width_multiple: 1.0 # layer channel multiple
anchors:
  - [10,13, 16,30, 33,23] # P3/8
  - [30,61, 62,45, 59,119] # P4/16
  - [116,90, 156,198, 373,326] # P5/32

# YOLOv5 v6.0 backbone
backbone:
  # [from, number, module, args]
  [[-1, 1, Conv, [64, 6, 2, 2]], # 0-P1/2
   [-1, 1, Conv, [128, 3, 2]], # 1-P2/4
   [-1, 3, C3, [128]],
   [-1, 1, Conv, [256, 3, 2]], # 3-P3/8
   [-1, 6, C3, [256]],
   [-1, 1, Conv, [512, 3, 2]], # 5-P4/16
   [-1, 9, C3, [512]],
   [-1, 1, Conv, [1024, 3, 2]], # 7-P5/32
   [-1, 3, C3, [1024]],
   [-1, 1, SPPF, [1024, 5]], # 9
  ]
```

```
# YOLOv5 v6.0 head
head:
  [[-1, 1, Conv, [512, 1, 1]],
   [-1, 1, nn.Upsample, [None, 2, 'nearest']],
   [[-1, 6], 1, Concat, [1]], # cat backbone P4
   [-1, 3, C3, [512, False]], # 13

   [-1, 1, Conv, [256, 1, 1]],
   [-1, 1, nn.Upsample, [None, 2, 'nearest']],
   [[-1, 4], 1, Concat, [1]], # cat backbone P3
   [-1, 3, C3, [256, False]], # 17 (P3/8-small)

   [-1, 1, Conv, [256, 3, 2]],
   [[-1, 14], 1, Concat, [1]], # cat head P4
   [-1, 3, C3, [512, False]], # 20 (P4/16-medium)

   [-1, 1, Conv, [512, 3, 2]],
   [[-1, 10], 1, Concat, [1]], # cat head P5
   [-1, 3, C3, [1024, False]], # 23 (P5/32-large)

   [[17, 20, 23], 1, Detect, [nc, anchors]], # Detect (P3, P4, P5)
  ]
```

1 신경망 시각화

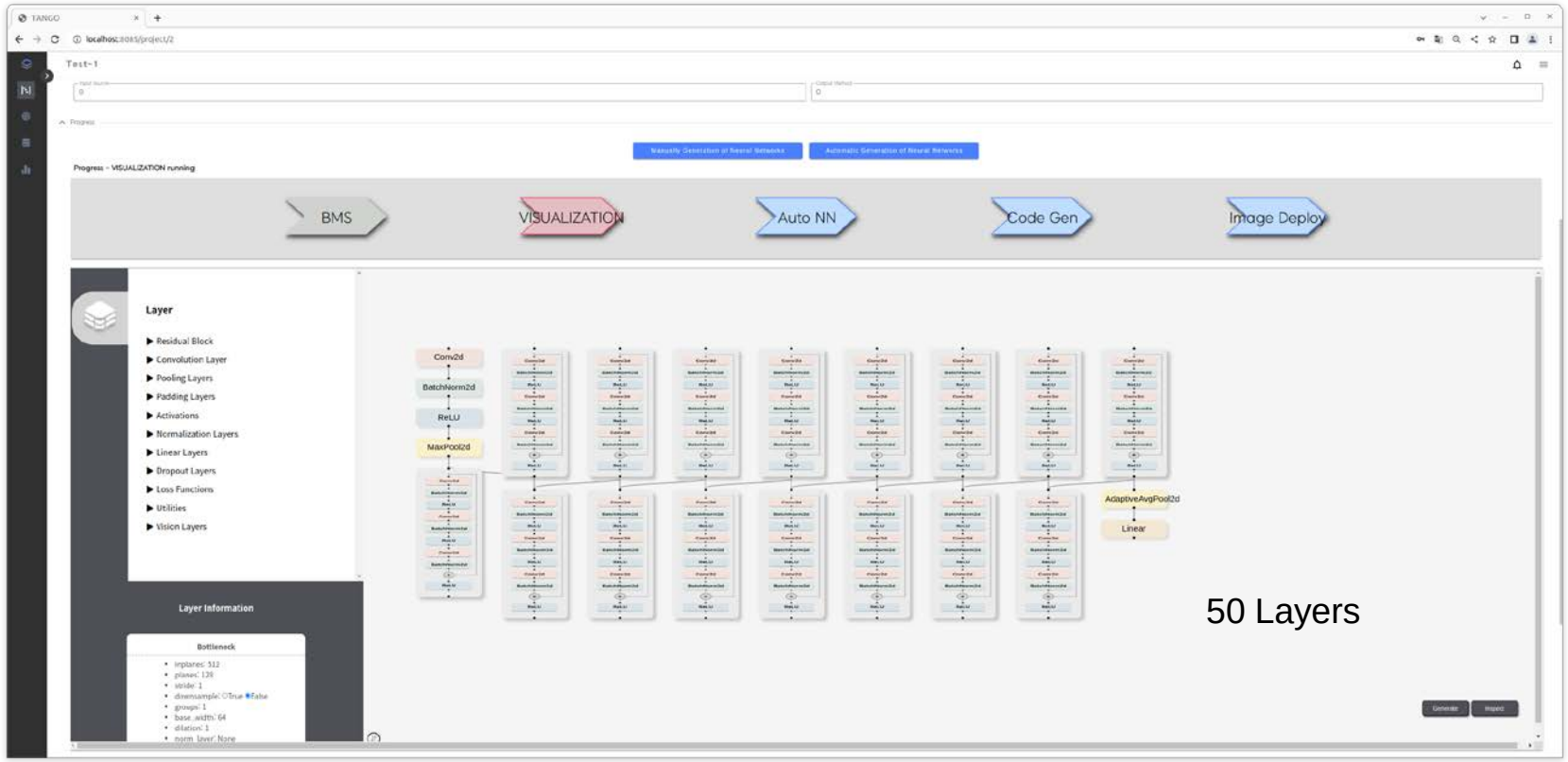
Yaml 생성 2

신경망 편집/수정



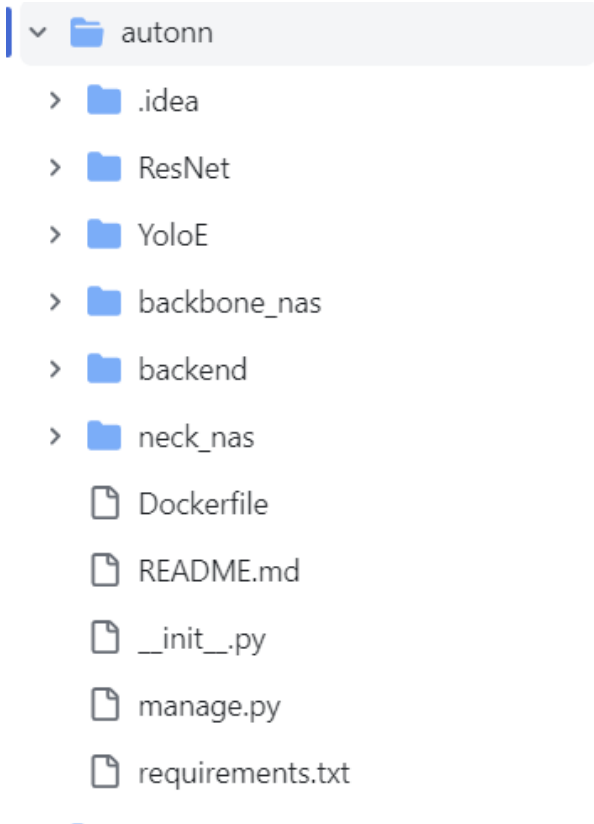
추천된 신경망 모델에 대해서 복잡도 증감 및 파라미터 세부 튜닝 기능 제공

- **신경망 시각화 기능**
  - 신경망을 파악할 수 있도록 모듈화 및 시각화
  - 신경망 복잡도 증감 혹은 내부 파라미터 세부 수정



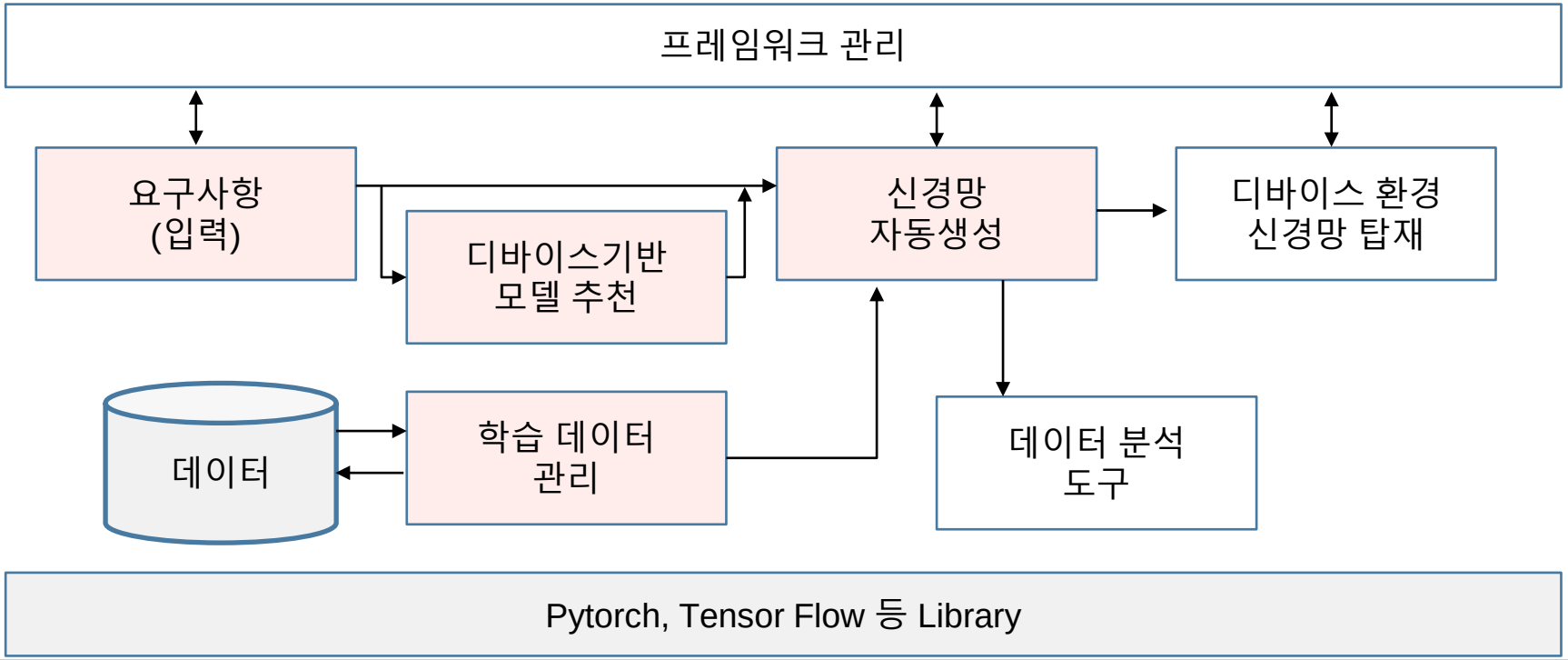
신경망 모델 추천에 따른 신경망 학습 및 최적 성능 지원

- **객체 탐지 및 이미지 분류를 위한 신경망 개발**
  - (1차) Yolov5를 이용한 백본망과 넥망의 독립적인 성능 향상 전략
  - (2차) Yolov7를 이용한 종합적인 성능 향상 및 On-devic에 대한 NAS 기능 강화
  - (2차) ResNet 신경망을 이용한 의류 이미지 분류 지원
- **신경망 주요 내용**
  - ResNet : 이미지 분류를 위해 34, 50, 101, 152 복잡도 지원
  - YoloE :
    - PC, Cloud를 위한 Yolov7 기본 기능 제공
    - On Device를 위한 NAS 기능 제공(스마트 폰)
    - 성능 향상을 위한 3차원 Neck 기능 제공



(고려대상) 효율적인 신경망 모델 추천과 다양한 신경망 생성 기능 지원

- **상호 유기적인 연동에 따른 모듈의 기능 추가**
  - **산업 데이터 및 디바이스에 적용하기 위한 전이학습 강화 (입력 사항 수정 및 학습 기능 강화)**
  - **신경망 모델 추천의 세밀화 (다양한 학습 결과 저장 및 데이터 이용)**
  - **신경망 모델 추천의 세밀화 (다양한 학습 결과 저장 및 데이터 이용)**



3. 향후 개발 내용 – 신경망 모델 추천

(고려대상) 다양한 디바이스와 응용에 대한 신경망 모델의 세밀한 추천

▪ 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

AI가속 GPU

성능

Processor	CPU Cores	AI Accelerator	Year	CPU Q AI Score	CPU F AI Score	QUANT Score	QUANT Accuracy	FP16 Score	FP16 Accuracy	FP32 Score	FP PAR Score	Accuracy	AI Score
Unisoc Tiger T710	4x2 GHz Cortex-A75 & 4x1.8 GHz Cortex-A55	NPU / n.a.	2019	1400	2113	8255	96	15603	79	394	314	85	26039 [15]
Snapdragon 855 Plus	1x2.96 + 3x2.42 + 4x1.8 GHz Kryo 485	DSP (Hex. 690) + GPU (Adreno 640)	2019	2213	4132	6375	40	8960	30	1357	1078	34	24553 [14]
HiSilicon Kirin 810	2x2.27 GHz Cortex-A76 & 6x1.88 GHz A55	NPU / n.a.	2019	1085	2488	1895	80	17635	95	463	348	90	23944 [1]
Snapdragon 855	1x2.84 + 3x2.41 + 4x1.78 GHz Kryo 485	DSP (Hex. 690) + GPU (Adreno 640)	2018	2106	3431	5535	55	7342	37	1156	714	43	20564 [1]
Mediatek Helio P90	2x2.2 GHz Cortex-A75 & 6x2 GHz Cortex-A55	DSP x 2 + APU / n.a.	2018	1067	2036	8666	98	10120	94	140	69	96	20060 [1]
Exynos 9825 Octa	n.a.	GPU / n.a. + NPU x 2	2019	3598	2966	2077	60	91170	45	780	717	50	17514 [15]
HiSilicon Kirin 980	2x2.6 GHz + 2x1.92 GHz A76 & 4x1.8 GHz A55	NPU x 2 / n.a.	2018	1817	3447	222	60	10750	85	139	64	76	16684 [2]
Exynos 9820 Octa	2x2.7 GHz M4 & 2x2.3 GHz A75 & 4x2 GHz A55	GPU (Mali-G76 MP12) + NPU x 2	2018	1491	2645	1650	60	8367	45	744	701	50	15713 [15]
Snapdragon 845	4x2.8 GHz Kryo 385/G & 4x1.7 GHz Kryo 385/S	DSP (Hex. 685) + GPU (Adreno 630)	2018	1605	2172	2031	58	6327	38	916	683	45	13804 [1]
Snapdragon 730	2x2.2 GHz Kryo 470/G & 6x1.8 GHz Kryo 470/S	DSP (Hex. 688) + GPU (Adreno 618)	2019	1066	2082	2765	58	3681	37	460	374	44	10453 [1]

타겟 디바이스

추론 수행 후 리턴값

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							

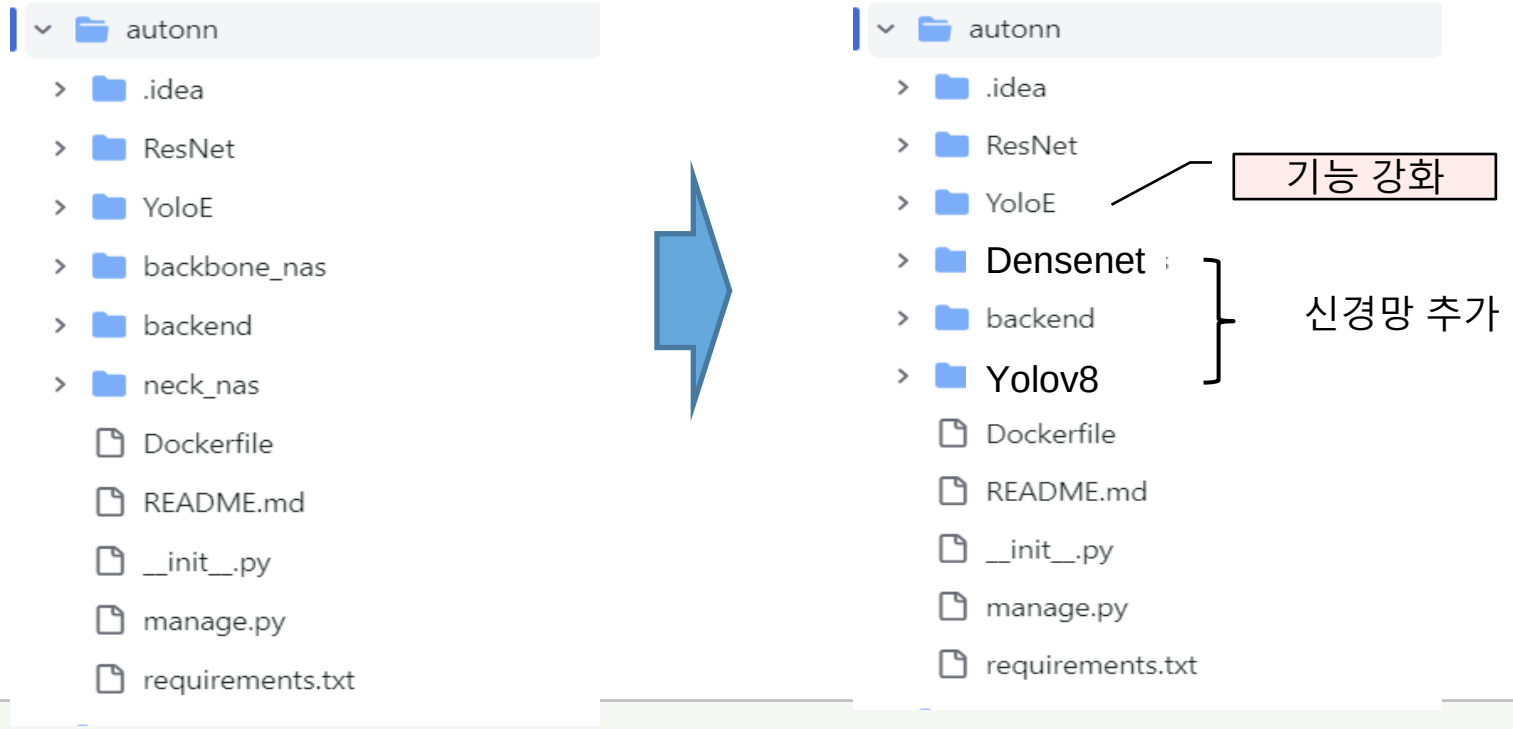
상세 정보는 NLP 이용

최적 신경망

3. 향후 개발 내용 – 신경망 생성

(고려대상) 산업 데이터에 대한 신경망 생성의 정확도 향상 및 확대 적용

- **타겟 적응형 신경망 자동 생성**
  - NAS **알고리즘의 확대 적용**
  - Bag of Freebies and Bag of Specials **알고리즘 다양화**
  - **신경망 모델의 지원 수 다양화** (Densenet)
  - **산업 적용을 위한 전이학습 기능 강화(데이터 레이블링과 연동 필요)**



3. 향후 개발 내용 – 입력 UI

(고려대상) 신경망 생성에서 추가적 기능을 제공하기 위한 UI 개선

- **학습의 세분화**
  - (기존) Scratch 학습 지원
  - (추가) 전이학습을 위한 Pretrained 모델을 사용하여 추가 학습하는 경우
- **디바이스 사양 검색 및 자동 분류 기능**
- **이미지 세그멘테이션에 대한** task type **적용 여부**

TANGO

감사합니다.

주관 ETRI 한국전자통신연구원 (TANGO) 주최 과학기술정보통신부 IITP 정보통신기획평가원

후원 UOS lablup weda tesla System (사)한국인공지능협회 Korea Artificial Intelligence Association SNUH 서울대학교병원 고려대학교 KOREA UNIVERSITY 홍익대학교 HONGIK UNIVERSITY 중앙대학교 Georgia Tech