인공지능 기술의 대중화(AI Democratization)를 위한

TANGO 커뮤니티 제1회 컨퍼런스

TANGO

신경망 백본 자동 탐색 기술

성명 이재성 소속 중앙대

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

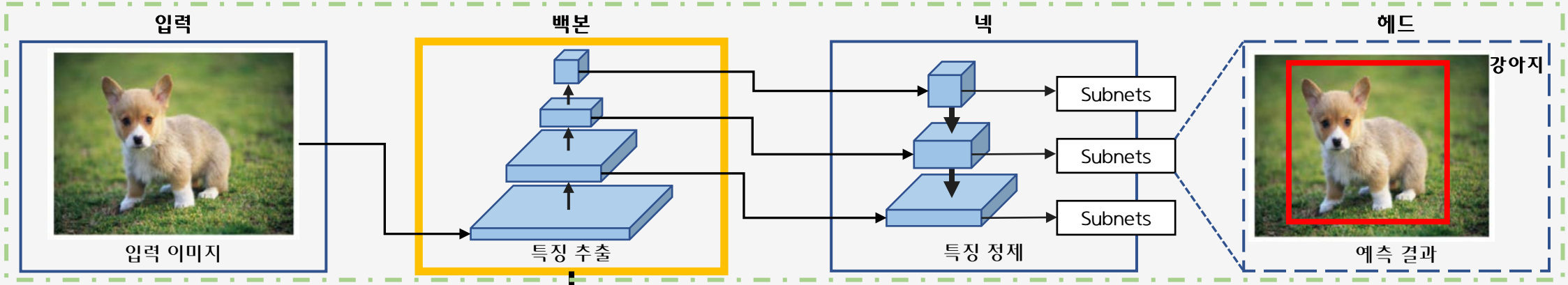
TANGO

목차

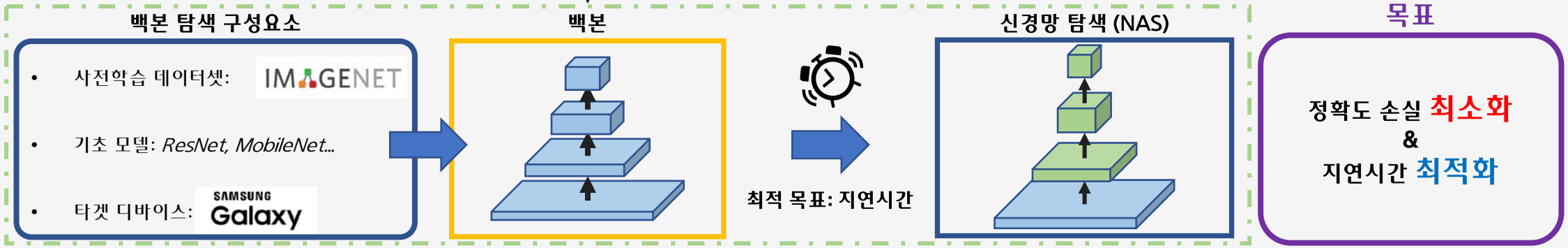
1. 기술 개요 - 신경망 백본 자동 탐색

연구 개요 및 범위

- 객체 검출



- 디바이스 최적 모델 탐색

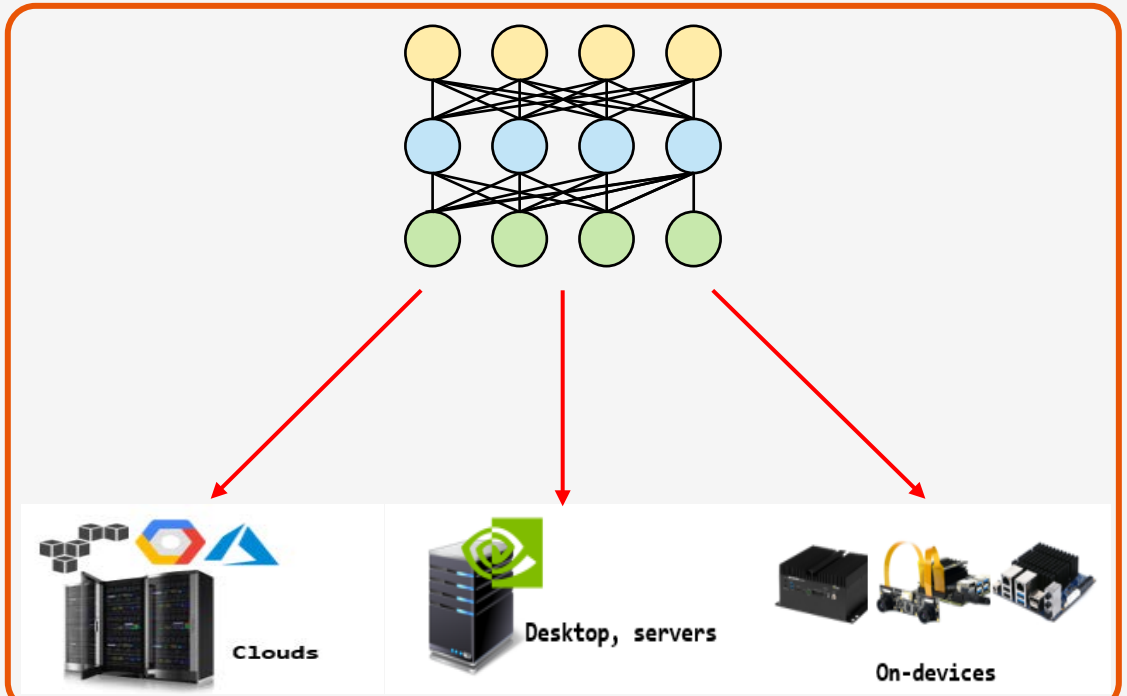


1. 기술 개요 – 신경망 백본 자동 탐색

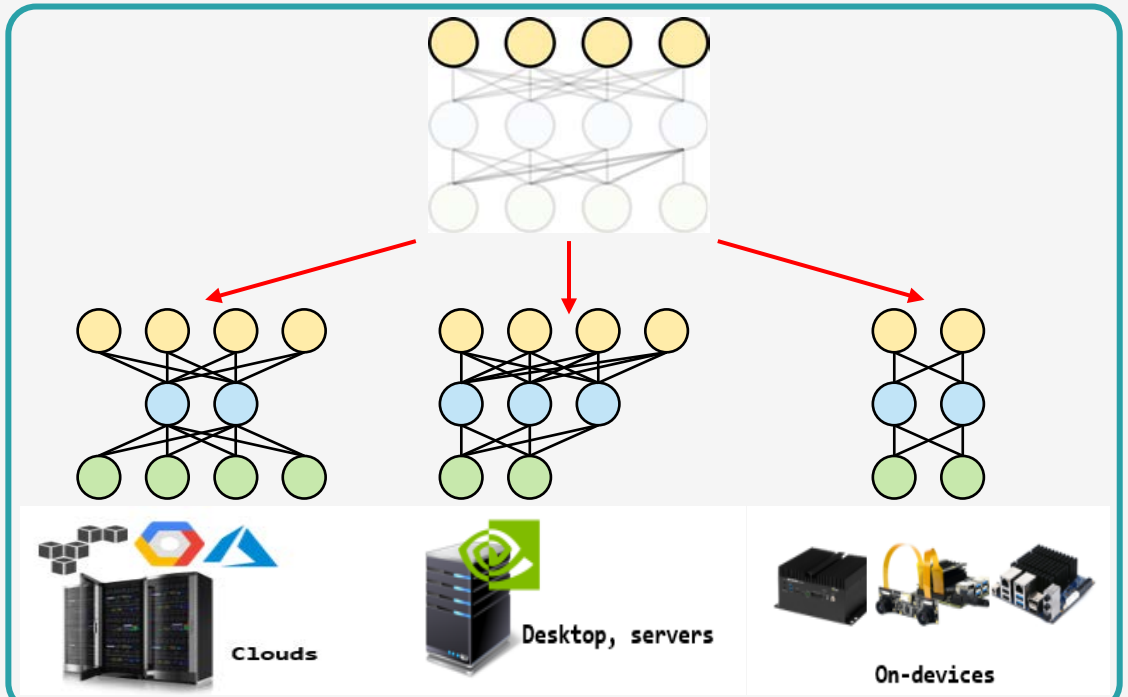
신경망 백본 자동 탐색 (BackboneNAS)

- 타겟 디바이스에 적합한 신경망 백본 자동 탐색

일반화된 모델



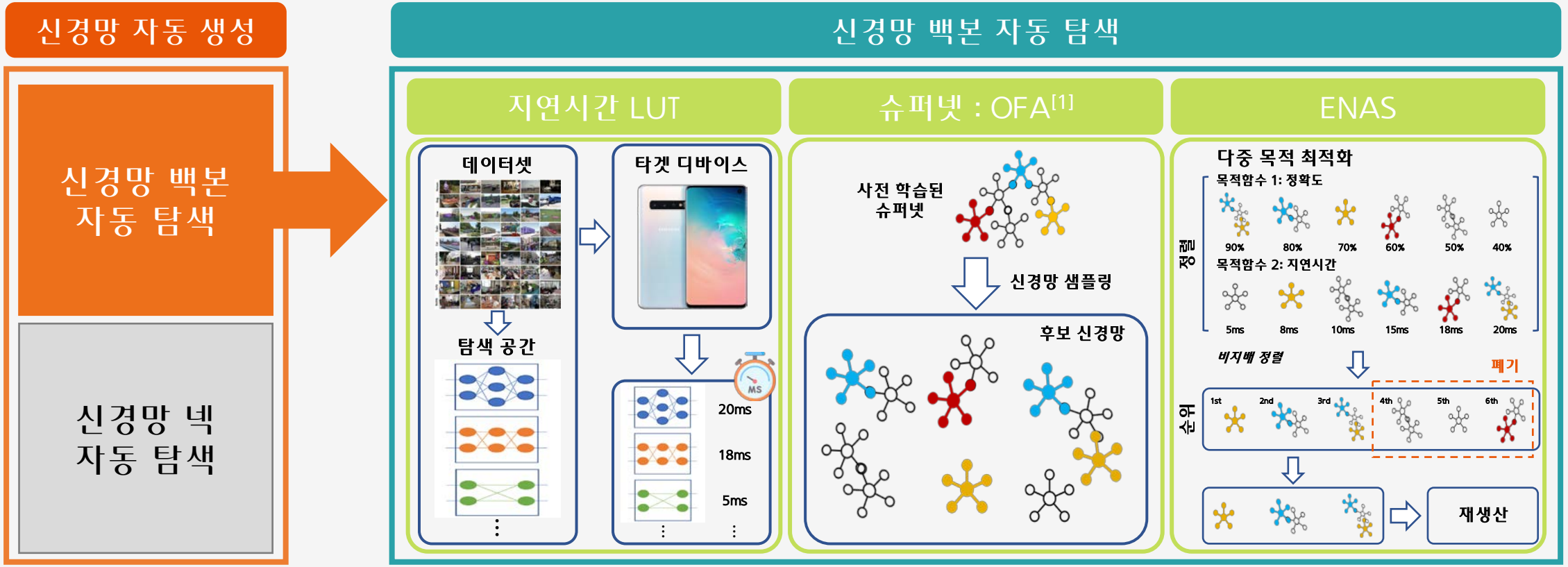
하드웨어에 특화된 모델



1. 기술 개요 - 신경망 백본 자동 탐색 프레임워크

신경망 백본 자동 탐색 프레임워크

- 지연시간 Lookup Table(LUT), 슈퍼넷, 진화 알고리즘 기반 탐색(ENAS)을 이용한 신경망 백본 자동 탐색



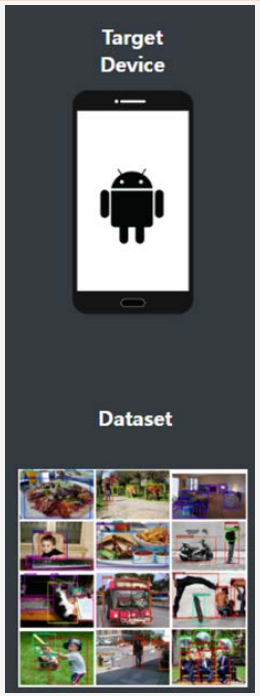
[1] Cai, Han, et al. "Once for All: Train One Network and Specialize it for Efficient Deployment." (ICLR). 2020.

1. 기술 개요 - 신경망 백본 자동 탐색 프레임워크

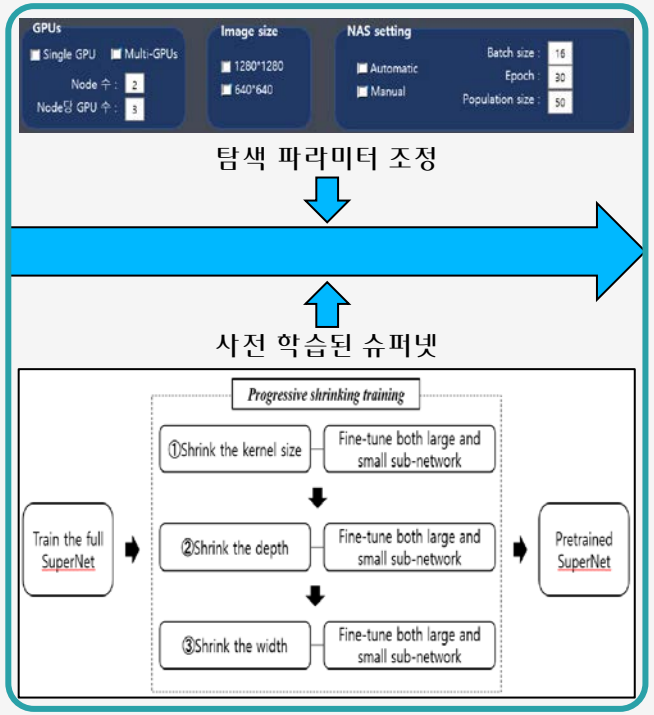
신경망 백본 자동 탐색 프레임워크

- 신경망 백본 자동 탐색을 위한 워크플로우

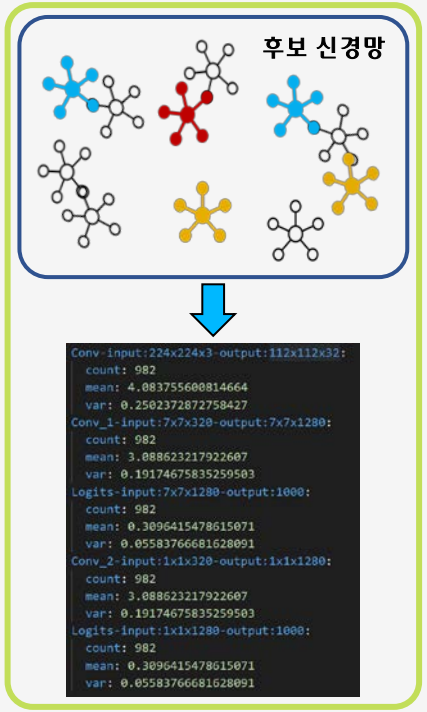
사용자 요구사항



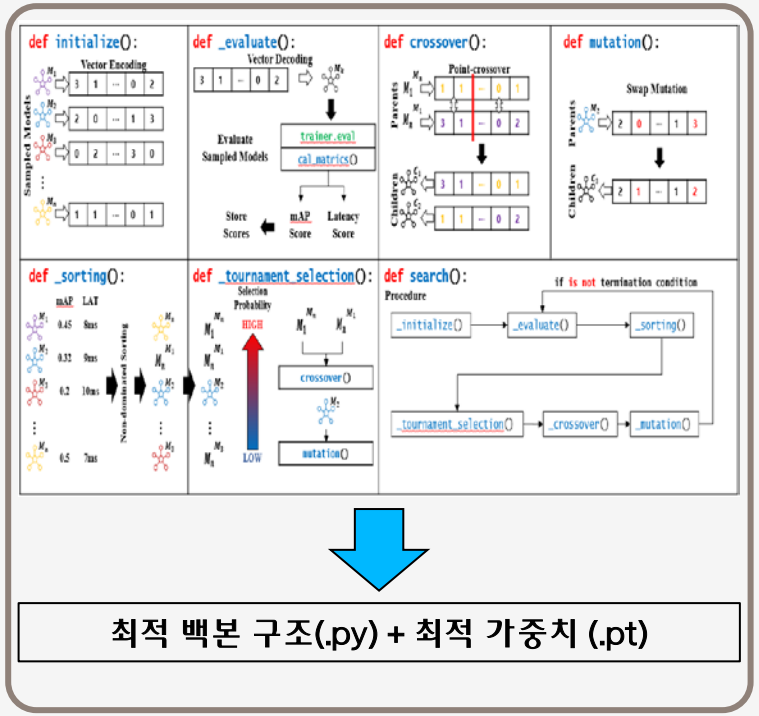
파라미터 설정



지연시간 LUT



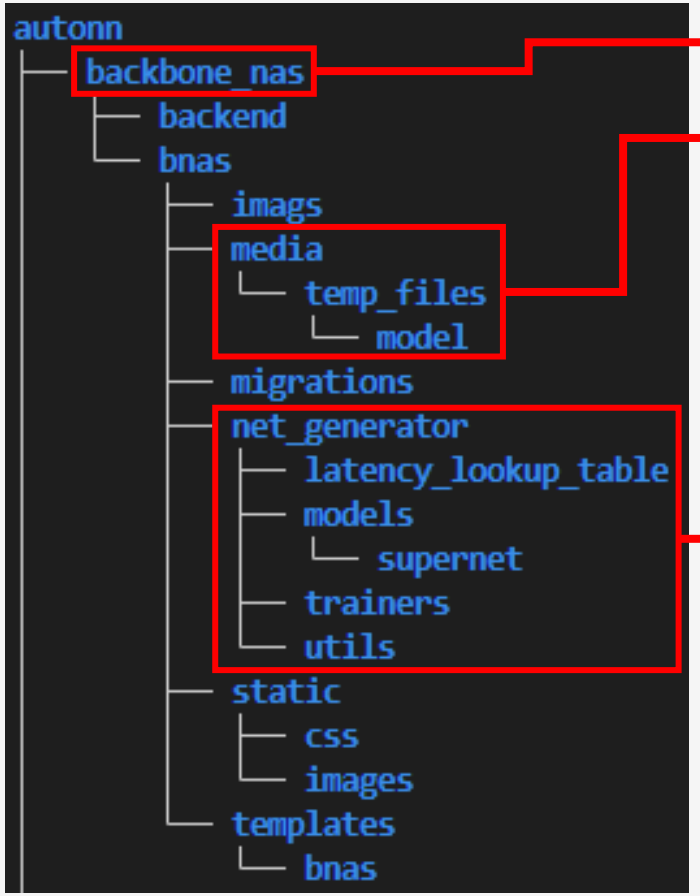
ENAS



1. 기술 개요 - 디렉토리 트리

디렉토리 트리

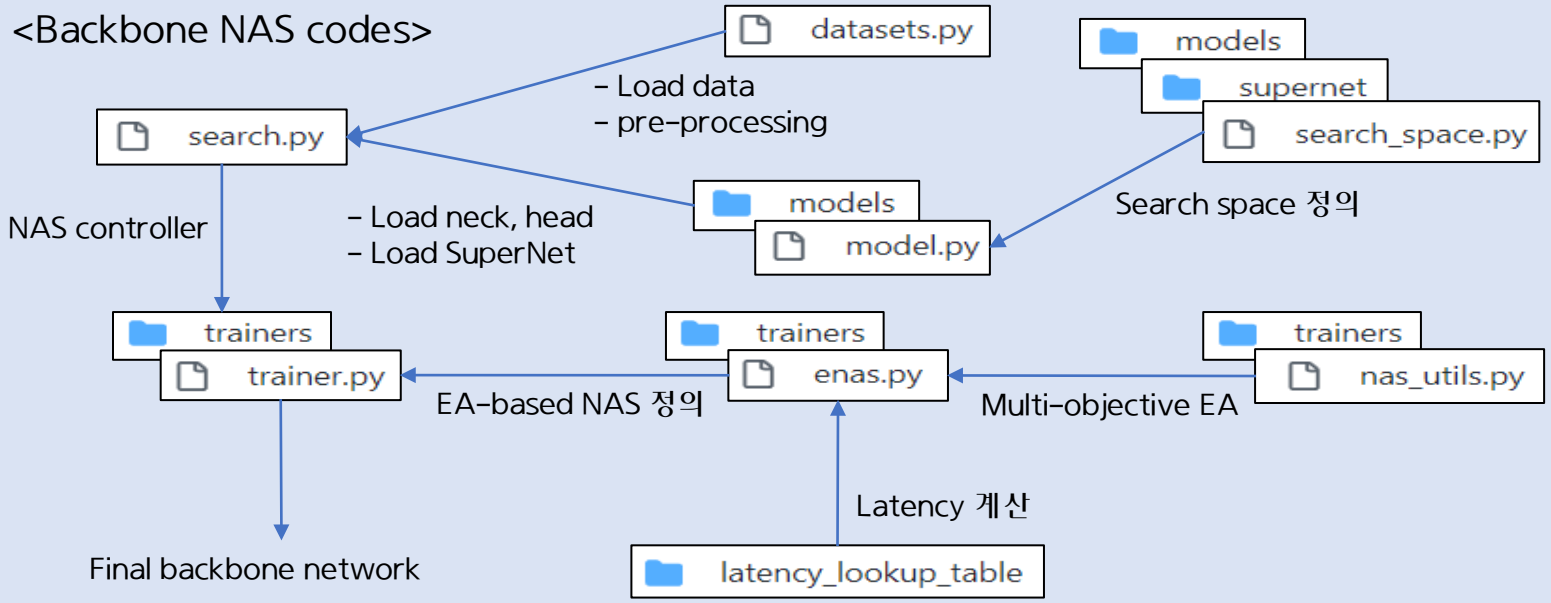
- 전문가용 활용을 위한 디렉토리 트리



- URL : http://localhost:8087/

- Base model: 슈퍼넷, yolov5[2] 저장
- Best model 임시 저장, final model은 공유 폴더에 저장

<Backbone NAS codes>



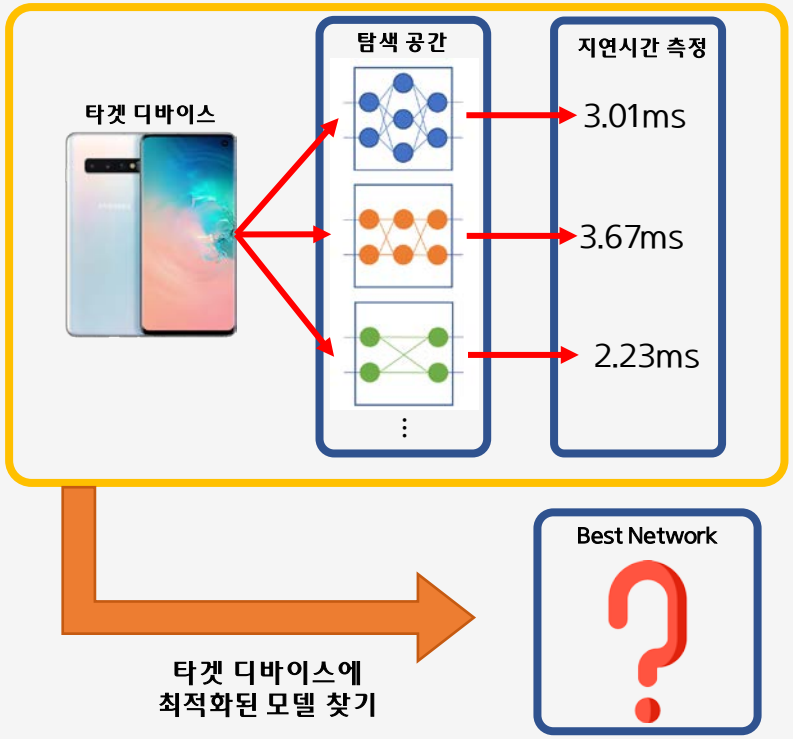
[2] https://github.com/ultralytics/yolov5

2. 개발 내용 – 지연시간 LUT

지연시간(Latency) LUT

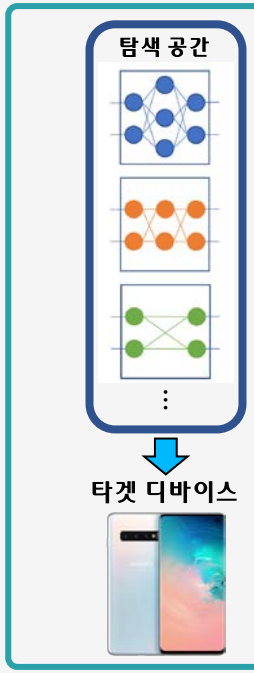
- 타겟 디바이스에 최적화된 모델 설계
- 타겟 디바이스 자원을 활용한 지연시간 LUT 생성

LUT 기반 모델 탐색 플로우



사용자 요구사항

지연시간 측정



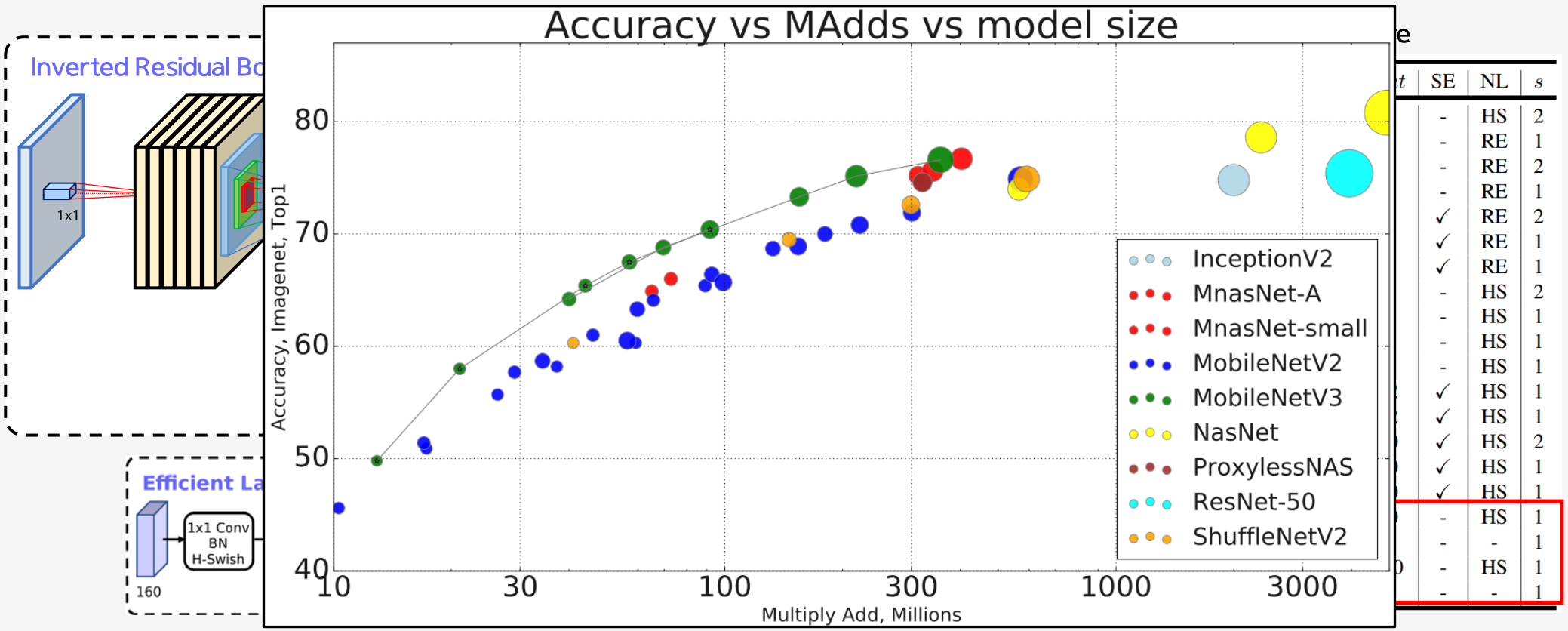
지연시간 LUT

탐색 블록	지연시간
Conv1	4.08ms
Expanded_Conv1	3.88ms
Expanded_Conv2	6.25ms
Expanded_Conv3	22.11ms
Expanded_Conv4	11.09ms
Expanded_Conv5	2.12ms
⋮	

2. 개발 내용 - 슈퍼넷 설계

슈퍼넷 설계 : 탐색 공간(Search space) 정의

- MobileNetV3[3] 기반의 Operator와 Architecture



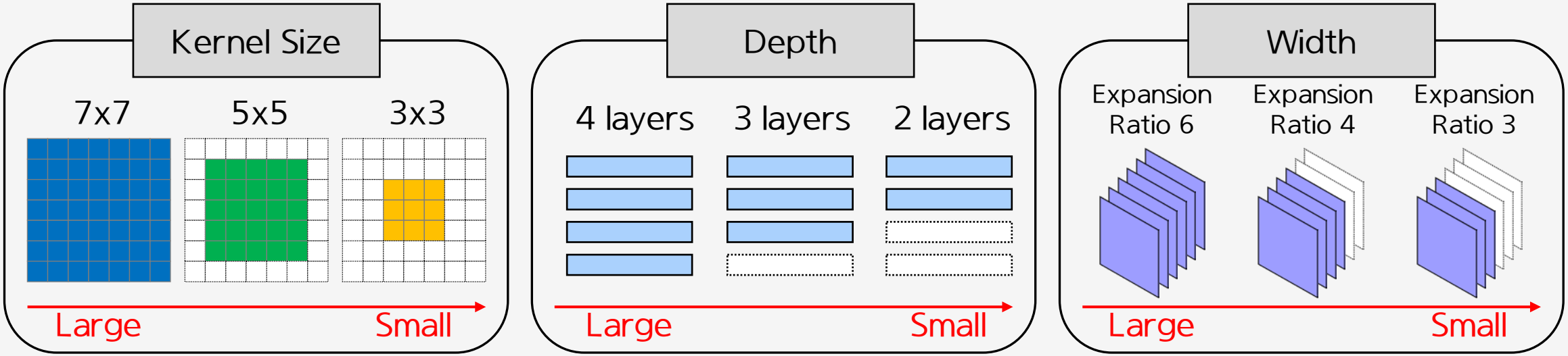
	SE	NL	s
	-	HS	2
	-	RE	1
	-	RE	2
	-	RE	1
	✓	RE	2
	✓	RE	1
	✓	RE	1
	-	HS	2
	-	HS	1
	-	HS	1
	-	HS	1
	✓	HS	1
	✓	HS	1
	✓	HS	2
	✓	HS	1
	✓	HS	1
	-	HS	1
	-	-	1
	-	HS	1
	-	-	1

[3] Howard, Andrew, et al. "Searching for mobilenetv3." (ICCV). 2019.

2. 개발 내용 - 슈퍼넷 설계

슈퍼넷 설계 : 탐색 공간(Search space) 정의

- 탐색 범위: Kernel Size, Width(Expansion Ratio), Depth(Number of layers)
  - ✓ 해당 파라미터들을 탐색함으로써 다양한 디바이스를 위한 후보 신경망 생성 가능



2. 개발 내용 – 슈퍼넷 설계

슈퍼넷 설계 : 탐색 공간(Search space) 정의

- 예시) 후보 신경망의 생성 과정

1) First layers (fixed)

   ✓ a conv layer

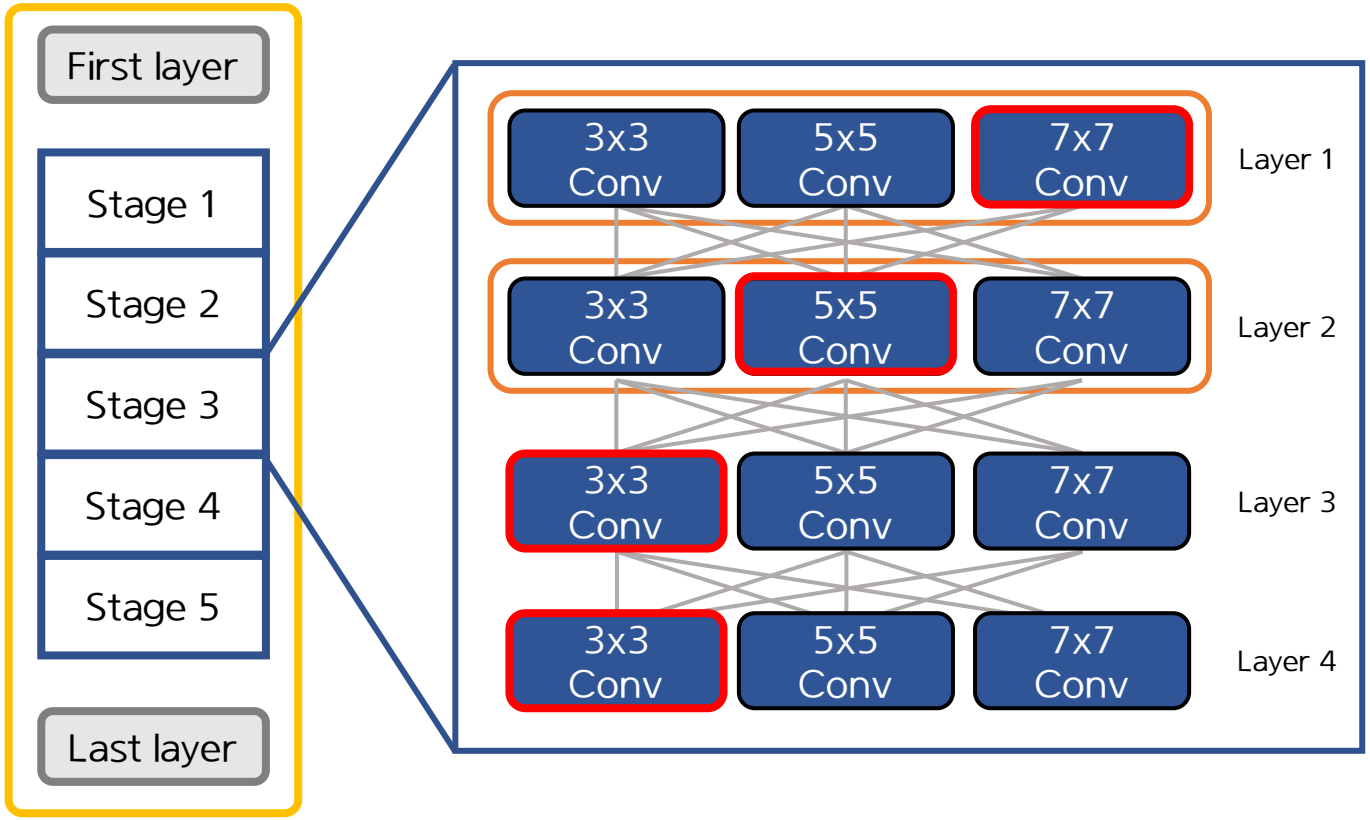
2) Search space
   (5 stages to be optimized)

   ✓ kernel size ∈ {3,5,7}

   ✓ depth ∈ {2,3,4}

   ✓ width ∈ {3,4,6}
   (expansion ratio)

3) Last layers (fixed)



2. 개발 내용 - 슈퍼넷 설계

슈퍼넷 설계 : 탐색 공간(Search space) 정의

- 예시) 후보 신경망의 생성 과정

1) First layers (fixed)

   ✓ a conv layer

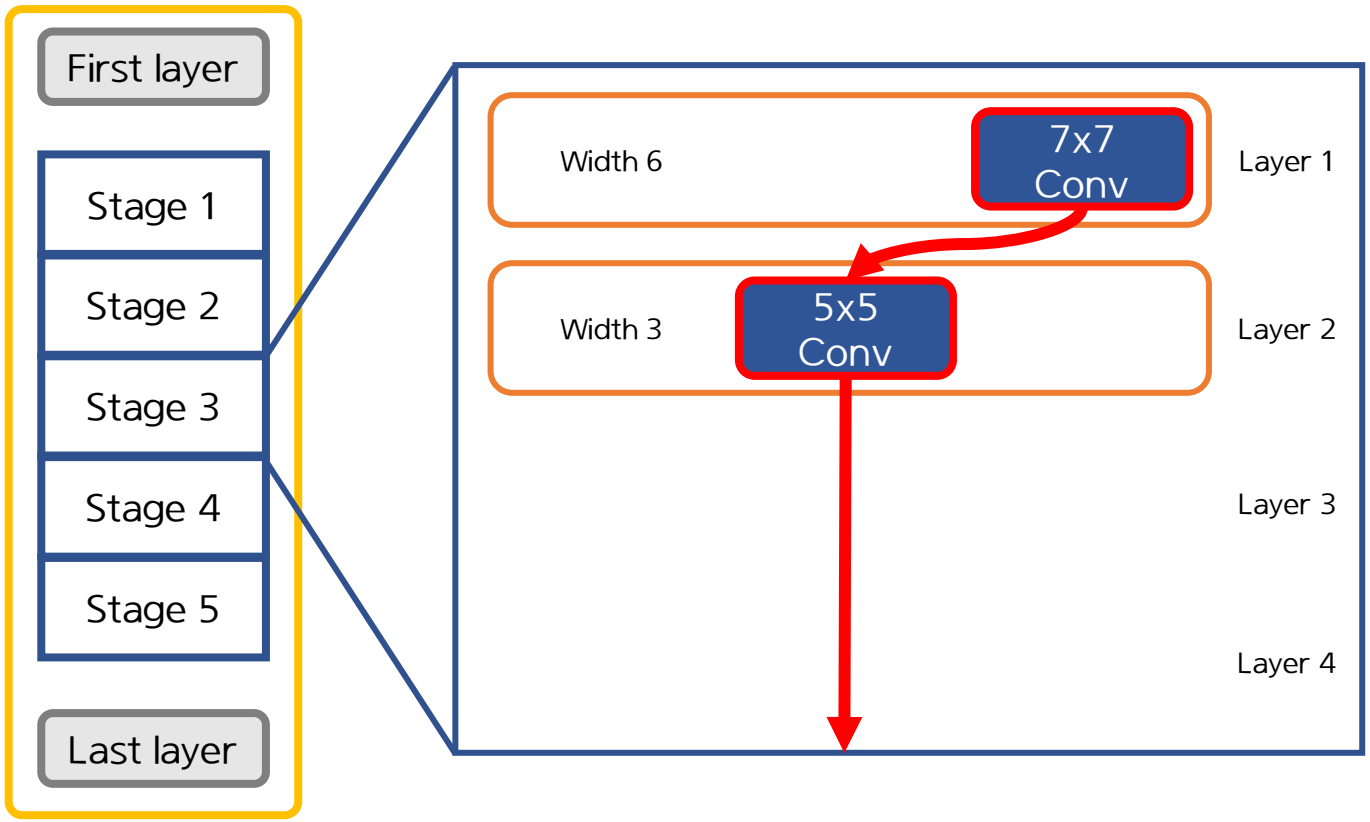
2) Search space
   (5 stages to be optimized)

   ✓ kernel size ∈ {3,5,7}

   ✓ depth ∈ {2,3,4}

   ✓ width ∈ {3,4,6}
   (expansion ratio)

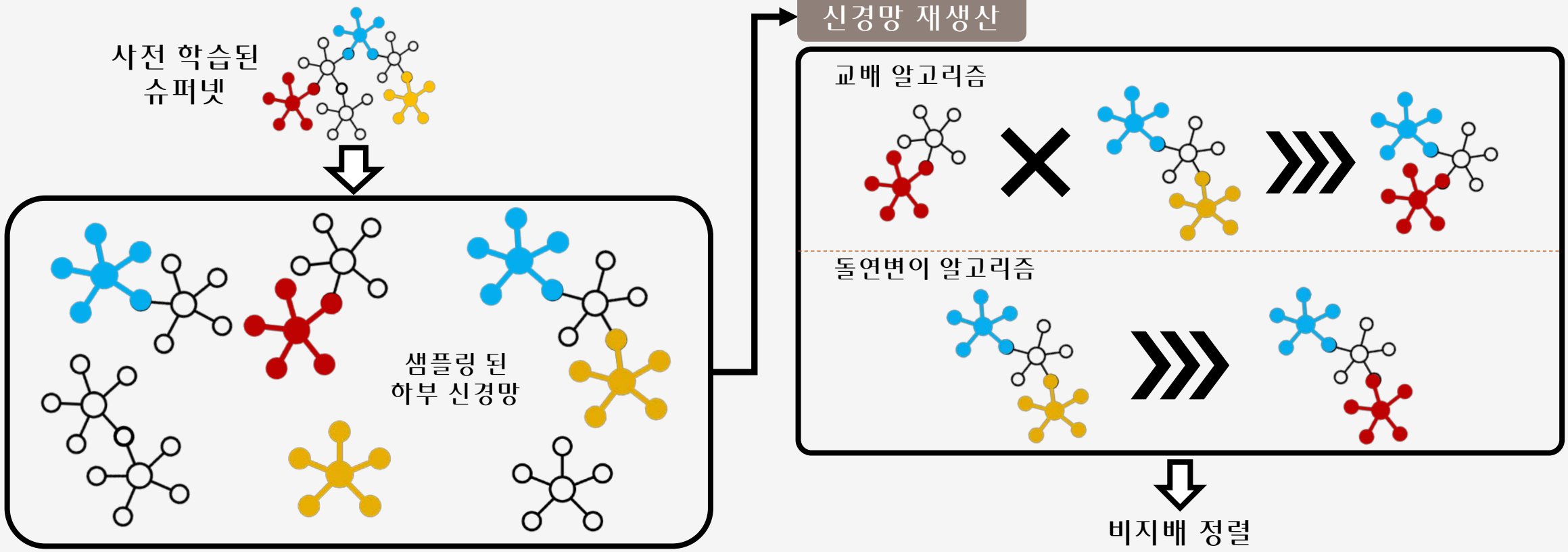
3) Last layers (fixed)



2. 개발 내용 – 진화 알고리즘 기반 탐색

슈퍼넷을 활용한 새로운 신경망 생성

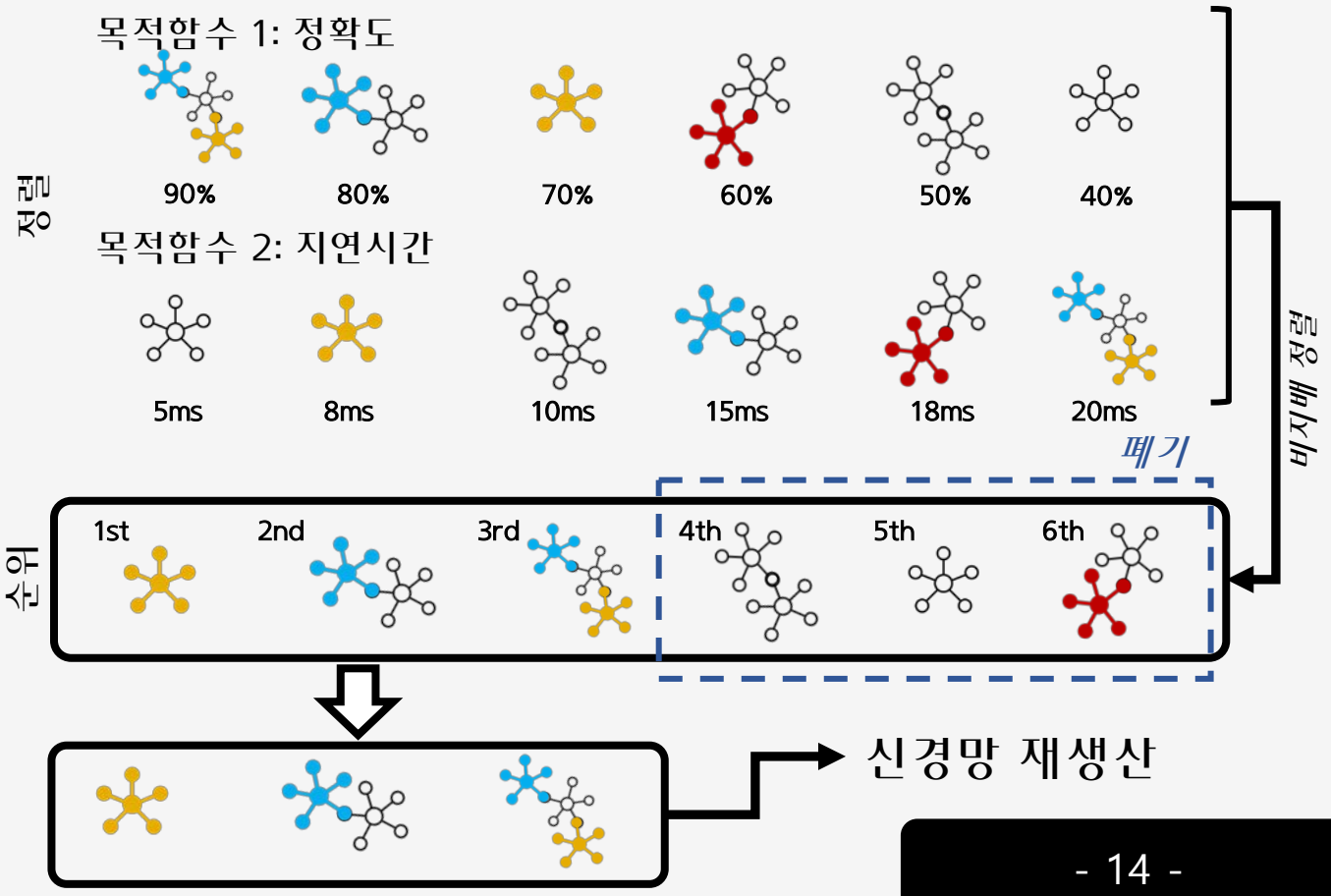
- 교배 및 돌연변이 알고리즘을 통한 하부 신경망 재생산



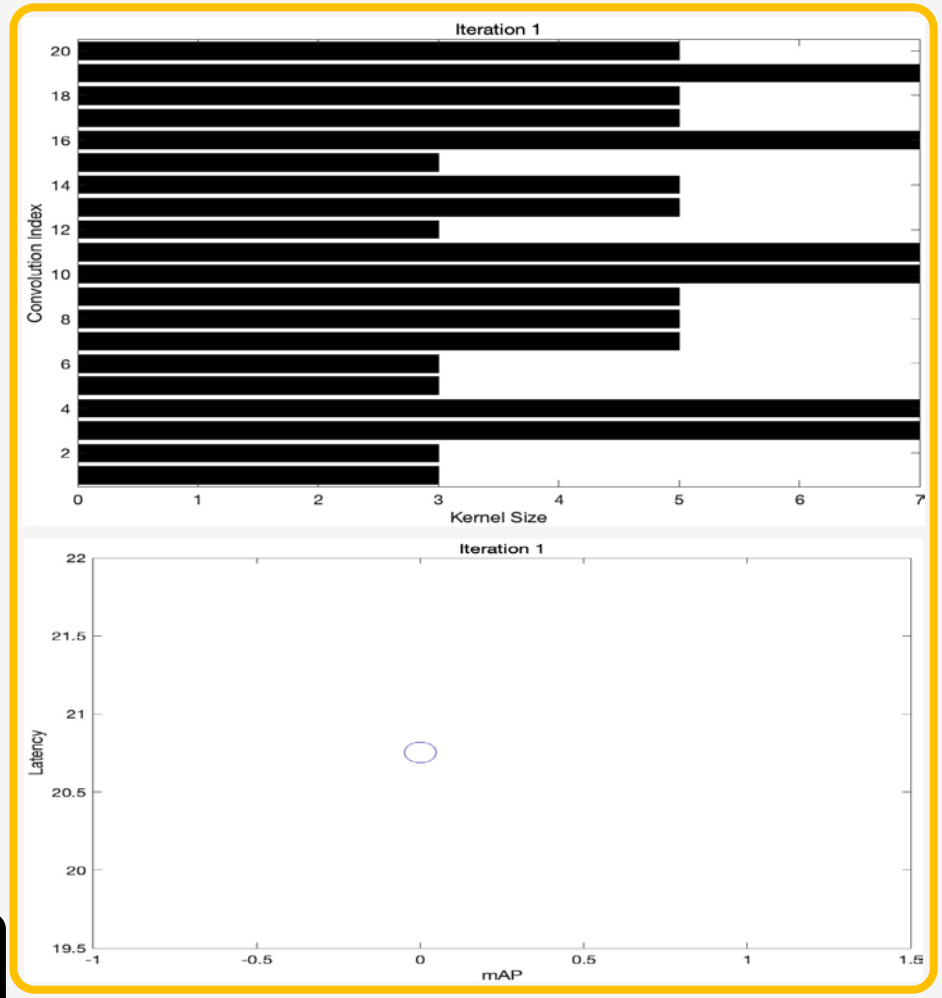
2. 개발 내용 – 진화 알고리즘 기반 탐색

정확도 – 지연시간 최적화를 위한 다중 최적화 알고리즘

- 예시) 비지배 정렬을 통한 하부 신경망 평가 및 정렬



신경망 구조 탐색



3. 향후 계획 – 관련 기술 비교

관련 기술들의 탐색 지원 범위

- Ours: 디바이스 최적 백본 신경망 지원, 정확도 개선이 요구

Model	Task	Search space				Objectives	
		Component	Kernel size	Depth	Width	Accuracy or mAP	Latency
ResNet	Cls	Backbone	–	–	–	✓	–
DetNet	Cls, Det	Backbone	–	–	–	✓	–
NAS-FPN	Det	Neck	–	–	–	✓	–
DetNAS	Cls, Det	Backbone	✓	–	–	✓	–
SP-NAS	Cls, Det	Backbone	–	✓	–	✓	–
Ours	Cls, Det	Backbone	✓	✓	✓	✓ (↓)	✓

3. 향후 계획 - 기술의 질적 향상

기술의 정량적 목표

평가 항목 (주요성능)		단위	전체 항목에서 차지하는 비중[2] (%)	세계 최고수준 보유국/보유기업	연구개발 전 국내 수준	연구개발 목표치		목표 설정 근거
				성능수준	성능수준	1단계(21~23)	2단계 (24~25)	
1.	자동생성 품질 (자동생성된 신경망 정확도[1])	%	15	미국/구글	-	53	55	o 구글 AutoML로 자동 생성된 객체 감지(detection) 신경망의 mAP 값으로 arXiv 2020년 7월 발표한 최신 데이터임 * 데이터 셋은 머신러닝을 위한 COCO DataSet으로 object detection, segmentation, keypoint detection 등에 많이 사용되고 있음
				53	-			
2.	타겟적응 품질 (신경망 연산 실행 Latency[2])	ms	20	미국/(페이스북-버컬리대)	-	28.1	25	o FBNets(Facebook-Berkeley-Nets) 모델을 사용하여 ImageNet Validation Set의 정확성과 CPU Latency를 측정한 결과로, 특정 타겟(삼성 갤럭시 S8) 기준으로 28.1ms의 Latency와 74.9%의 정확성(출처: https://arxiv.org/pdf/1812.03443.pdf)
				28.1	-			

감사합니다.

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