nota

On-device AI Startup

Nota Incorporated, which has a philosophy of using AI/ML
to make the world more convenient, started from
Korea Advanced Institute of Science and Technology(KAIST)

Tae-Ho Kim | CTO

www.nota.ai

Short Bio

B.S. Bio and Brain Engineering, KAIST

M.S. Electrical Engineering, KAIST

Research Intern, Universite de Montreal (P.I. Yoshua Bengio)

Research Intern, Chinese University of Hong Kong (P.I. Xiaogang Wang)

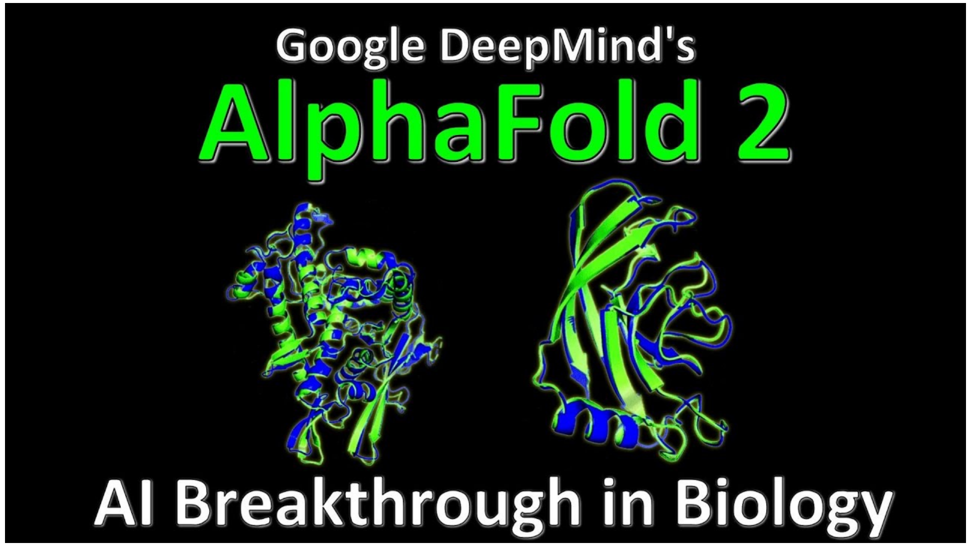
Senior Research Scientist, KAIST Institute

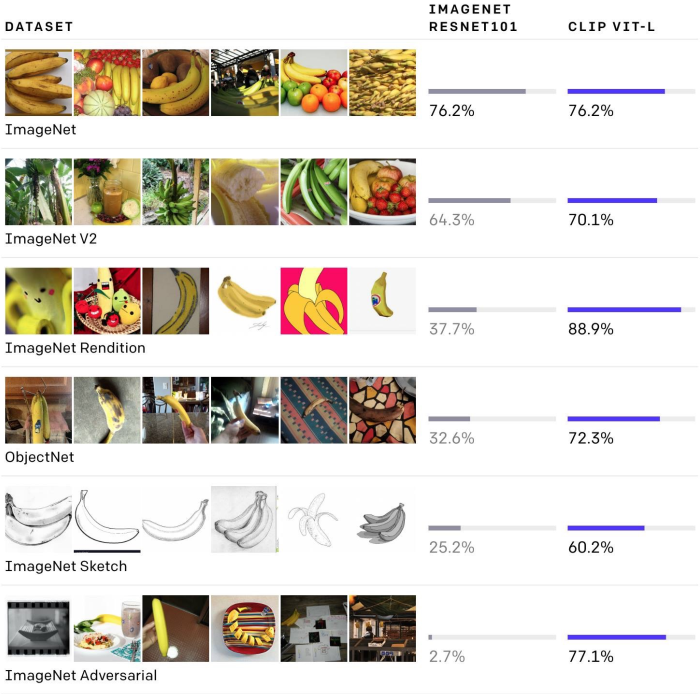
CTO / Co-Founder, Nota

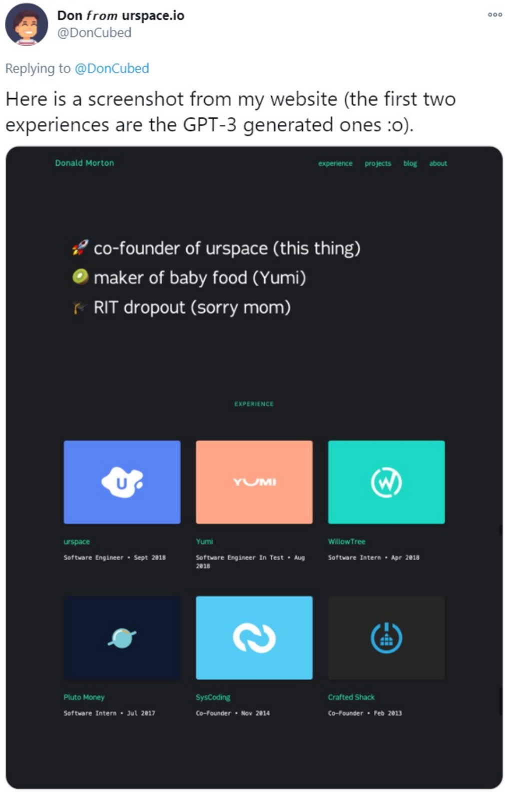
Research interests: Deep learning architecture itself, its application to CV, NLP, and Speech

Table of Contents

Achievements of Deep Learning



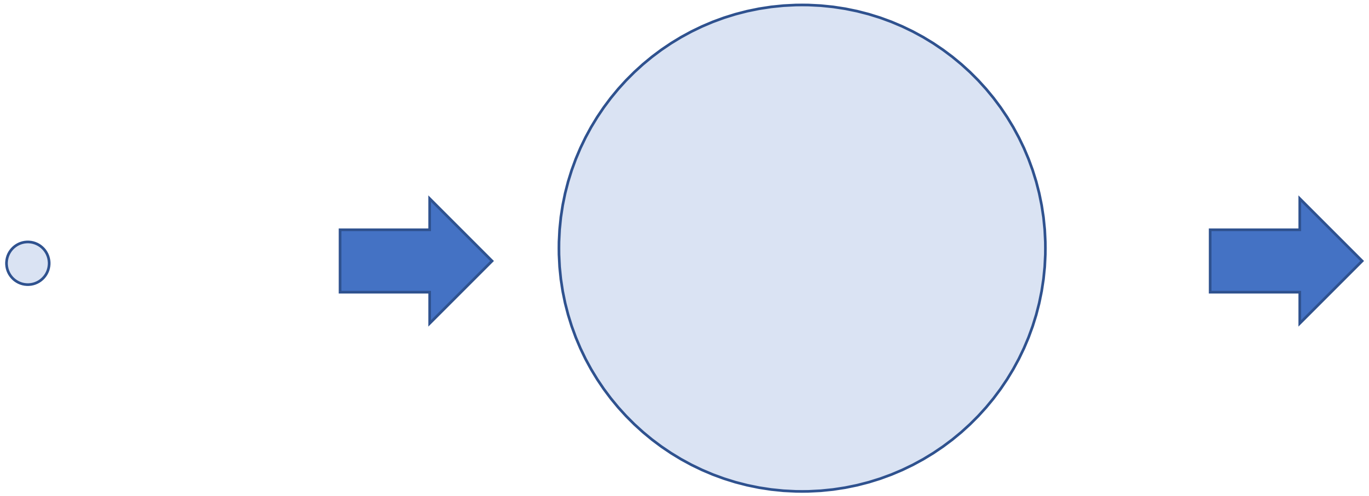




Future is coming

https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology
https://openai.com/blog/openai-api/
https://openai.com/blog/clip/

Deep Learning? How big?



Next?

GPT-2: 1.5B
Parameters

GPT-3: 175B
Parameters

It's going to be bigger…

Deep Learning? How big?

Training the Model

GPT-3 is trained using next word prediction, just the same as its GPT-2 predecessor. To train models of different sizes, the batch size is increased according to number of parameters, while the learning rate is decreased accordingly. For example, GPT-3 125M use batch size 0.5M and learning rate of 6.0×10^{-4}, where GPT-3 175B uses batch size 3.2M and learning rate of 0.6×10^{-4}.

We are waiting for OpenAI to reveal more details about the training infrastructure and model implementation. But to put things into perspective, GPT-3 175B model required 3.14E23 FLOPS of computing for training. Even at theoretical 28 TFLOPS for V100 and lowest 3 year reserved cloud pricing we could find, this will take 355 GPU-years and cost $4.6M for a single training run. Similarly, a single RTX 8000, assuming 15 TFLOPS, would take 665 years to run.

Time is not the only enemy. The 175 Billion parameters needs $175 \times 4 = 700GB$ memory to store in FP32 (each parameter needs 4 Bytes). This is one order of magnitude larger than the maximum memory in a single GPU (48 GB of Quadro RTX 8000). To train the larger models without running out of memory, the OpenAI team uses a mixture of model parallelism within each matrix multiply and model parallelism across the layers of the network. All models were trained on V100 GPU's on the part of a high-bandwidth cluster provided by Microsoft.

In fact, The size of SOTA language model increases by at least a factor of 10 every year: BERT-Large (2018) has 355M parameters, GPT-2 (early 2019) reaches 1.5B, T5 (late 2019) further streches to 11B, GPT-3 (mid-2020) finally gets to 175B. The progress of the sizes of language models clearly outpace the growth of GPU memory. This implies that for NLP, the days of "embarrassingly parallel" is coming to the end, and model parallelization is going to be indispensable for researching SOTA language models.

355 GPU-years

$4.6M

700GB

Super-huge!

https://lambdalabs.com/blog/demystifying-gpt-3/

Challenges of Cloud-based AI

High
Cost

Privacy
Concerns

Network Connectivity
Issues

Cloud-based AI becomes Unsustainable

AI Model Compression

Low
Cost

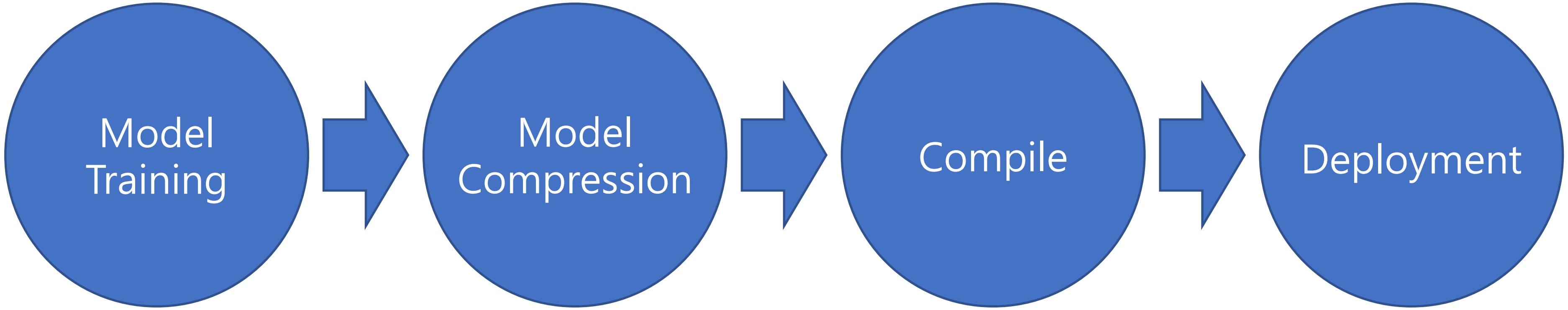
NO
Privacy Concerns

Stand-Alone
AI model

Solution : Nota's AI Model Compression Technology

nota
On-device AI

How can we use deep learning at the edge?



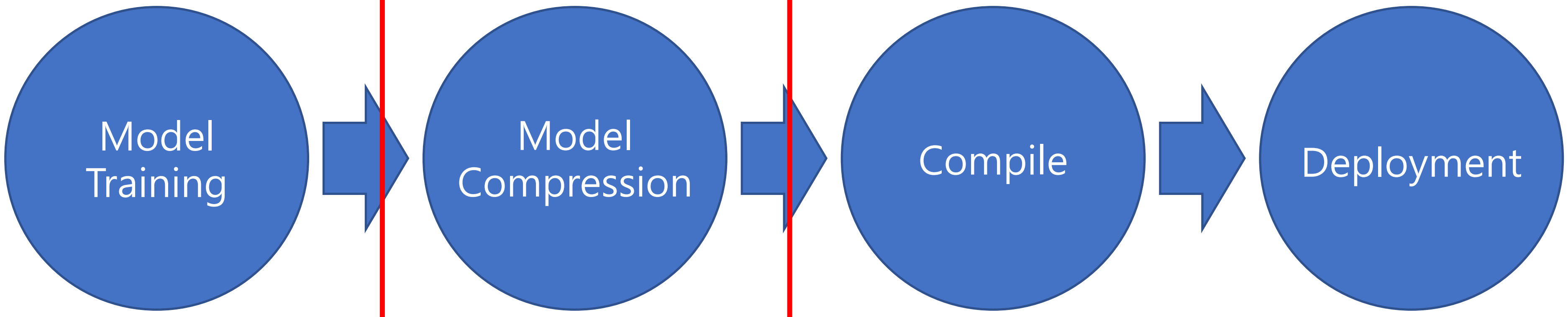
Model Training
- NAS (Neural Architecture Search)
- Popular models

Model Compression
- Filter Decomposition
- Knowledge Distillation
- Network Pruning
- Quantization
- Joint Approach

Compile
- Kernel Optimization
- Graph Optimization

Deployment
- HW improvement

How can we use deep learning at the edge?



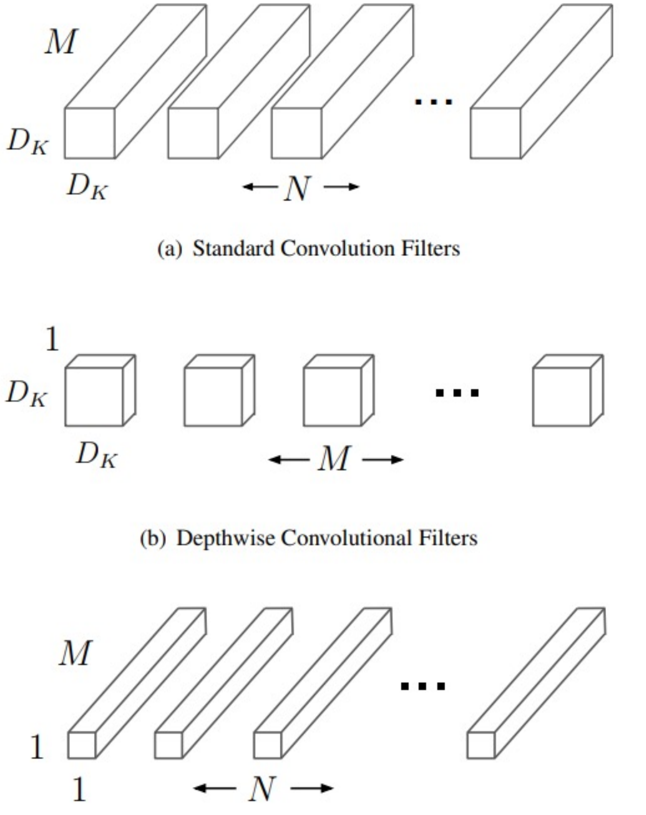
- NAS (Neural Architecture Search)
- Popular models

- Filter Decomposition
- Knowledge Distillation
- Network Pruning
- Quantization
- Joint Approach

- Kernel Optimization
- Graph Optimization

- HW improvement

Filter Decomposition



(c) 1×1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

Cost saving

$$\frac{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F}$$
$$= \frac{1}{N} + \frac{1}{D_K^2}$$

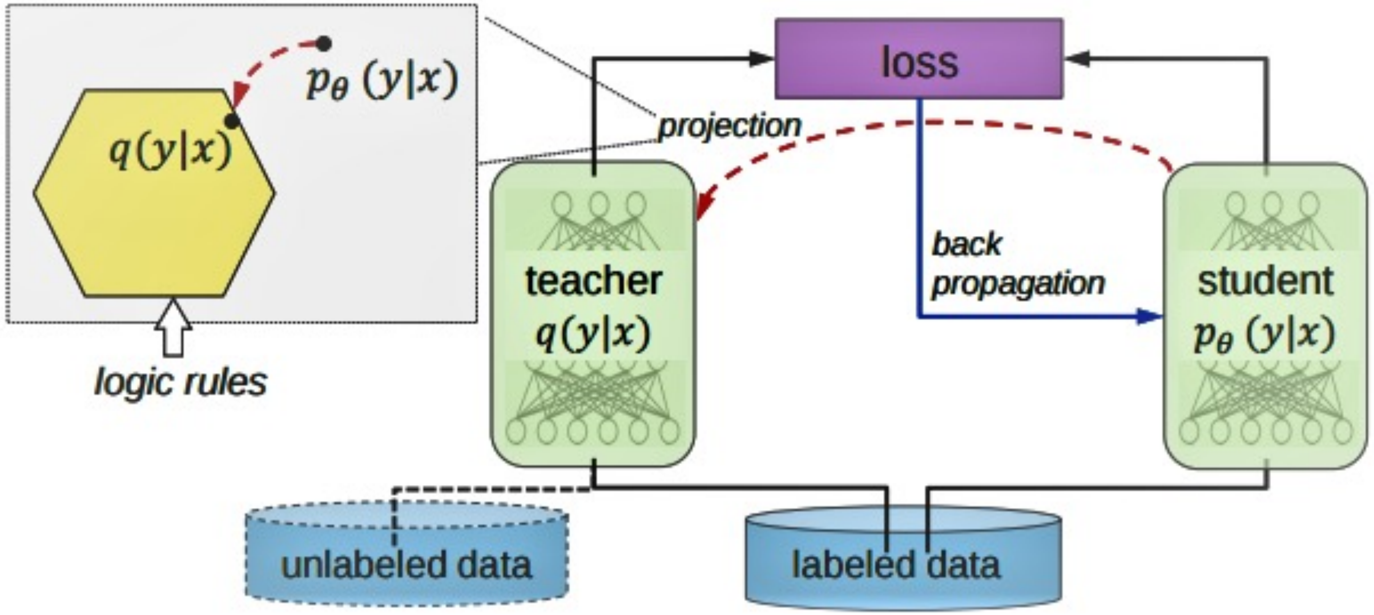
Types of filter decomposition

- Tucker Decomposition (Z. Zhong, 2019)
- **Depthwise Separable Convolution (A. Howard, 2017)**
- Network Decoupling (J. Guo, 2018)
- Truncated SVD
- …

MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, 2017

nota On-device AI

Knowledge Distillation



Types of Knowledge Distillation

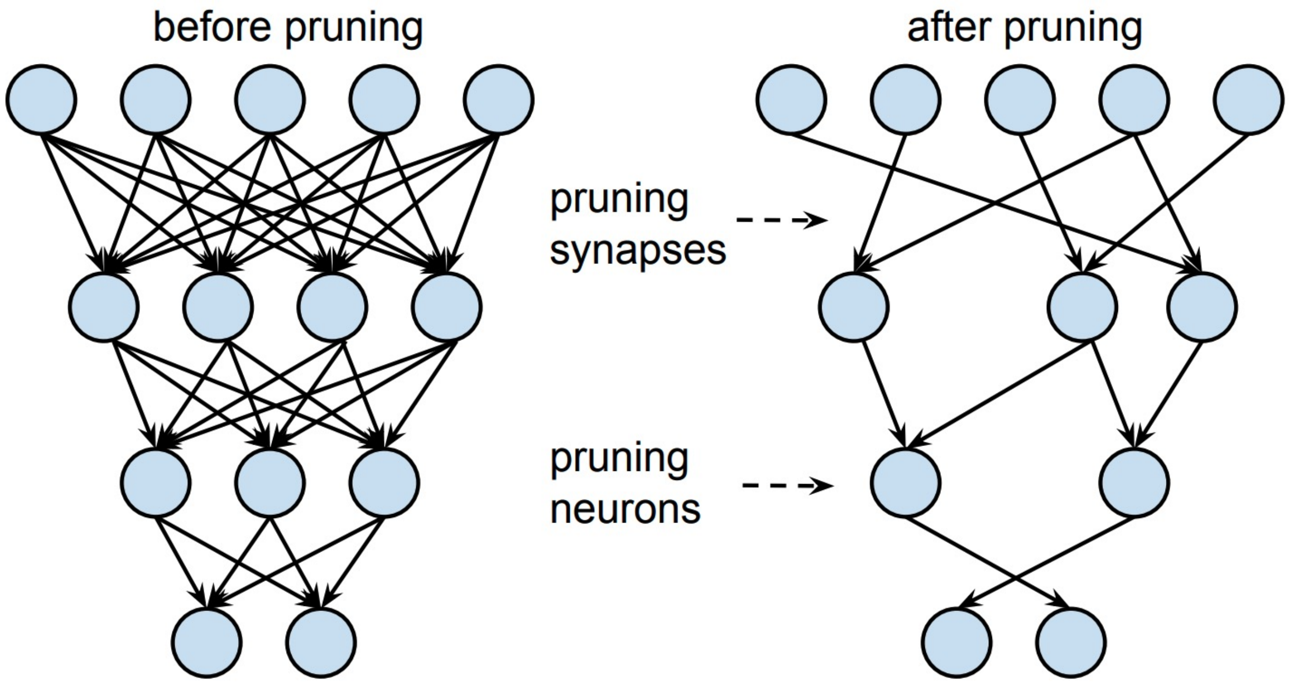
- Features Distillation
- Softlabel
- Attention Distillation

Techniques

- KD (G. Hinton, 2015)
- FitNets (A Romero, 2014)
- OverHaul KD (B Heo, 2019)
- Relational KD (W Park, 2019)
- …

Upadhyay, Ujjwal. 2018. "Knowledge Distillation." Medium, April 05.

Network Pruning



Weight? Filter? Channel?
- Structured Pruning
- Unstructured Pruning

Metrics
- L1 / L2 (S. Han, 2015)
- GM Pruning (Y He, 2018)
- BN Pruning (Y Liu, 2019)

Comparison scope
- Local Pruning
- Global Pruning

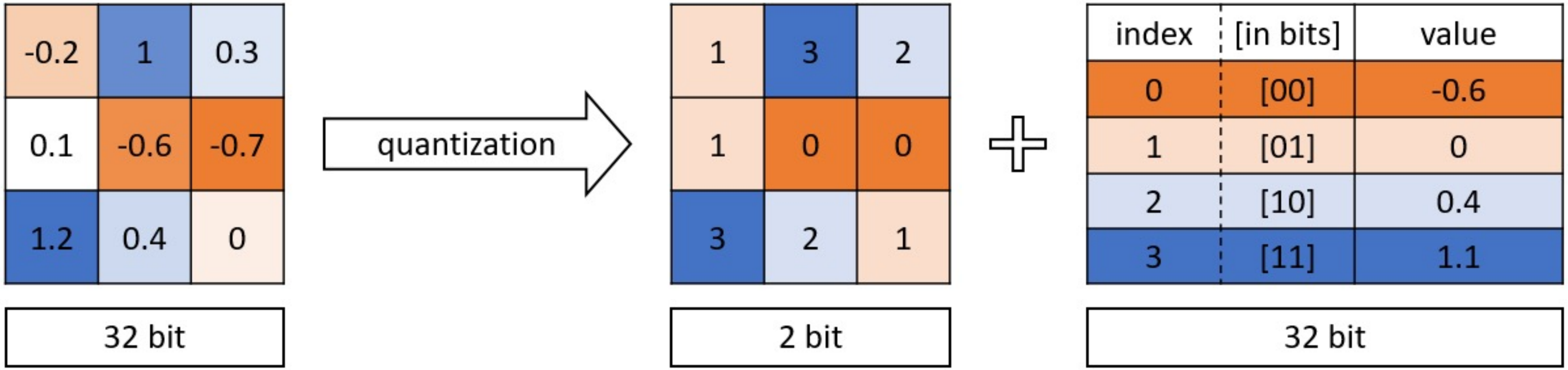
Slide credit: Song Han, 2015

nota
On-device AI

Quantization

- DoReFa (S Zhou, 2016)
- PACT (J Choi, 2018)
- QAT (Quantization Aware Training)
- PTQ (Post Training Quantization)
- …

- Bit? 0 or 1
- 2 bit variable can represent 4 numbers
- 32 bit variable can represent 2^32 numbers

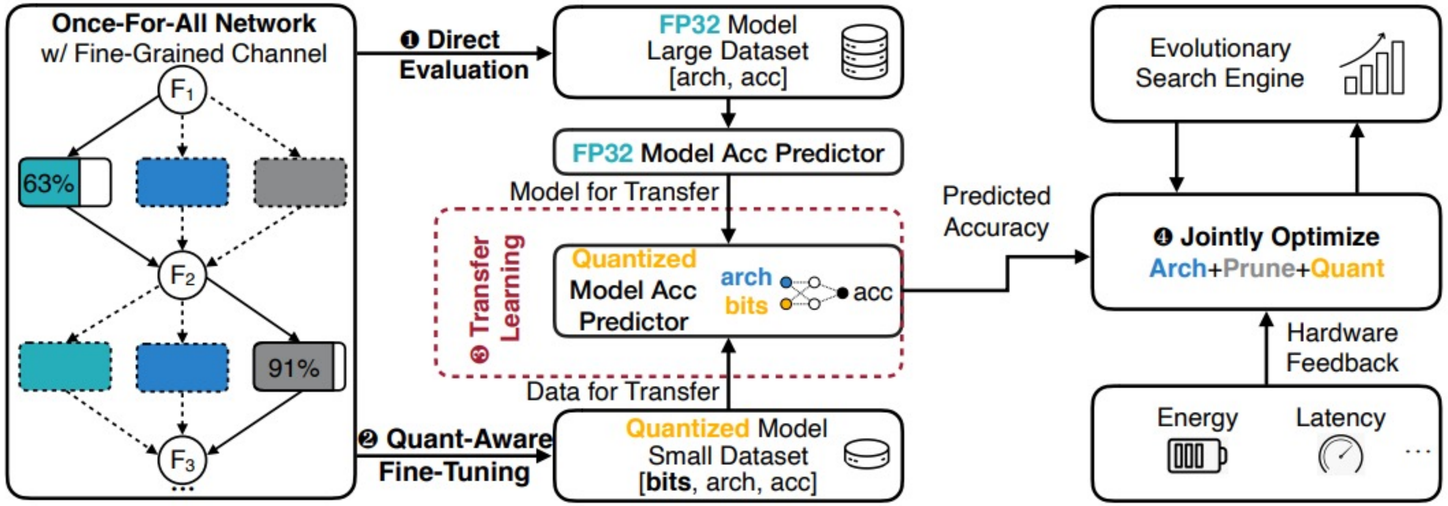


Slide credit: Song Han, 2015

nota On-device AI

Joint Approach

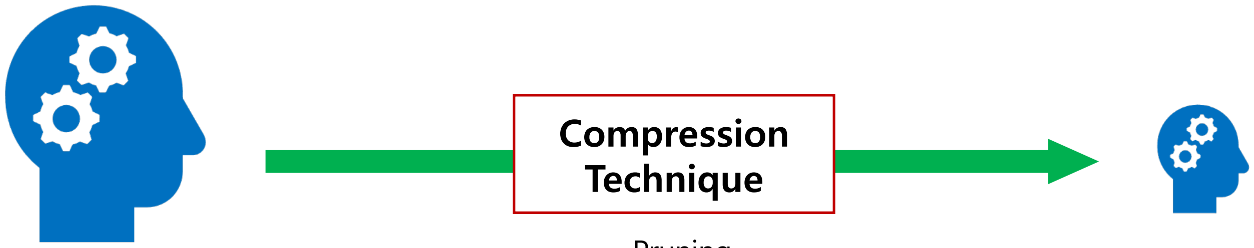
- Once-for-all: Considering pruning, KD, kernel size, and number of layers (ICLR 2020)
- APQ: Joint Search for Network Architecture, Pruning and Quantization Policy (CVPR 2020)



Then what?

- How can we compress the deep learning models with those techniques?
- Nota’s answer is coming…

Conventional AI model compression



Pretrained Model

Pruning

Quantization

Knowledge Distillation

NAS

Compressed Model

Problems of current network compression

- DL engineers **manually compress** the model
- Compression methods are developed **in different places and forms**
- **Hard to know** which compression method or combination to use
- Compression metric **does not fit** to practical metric

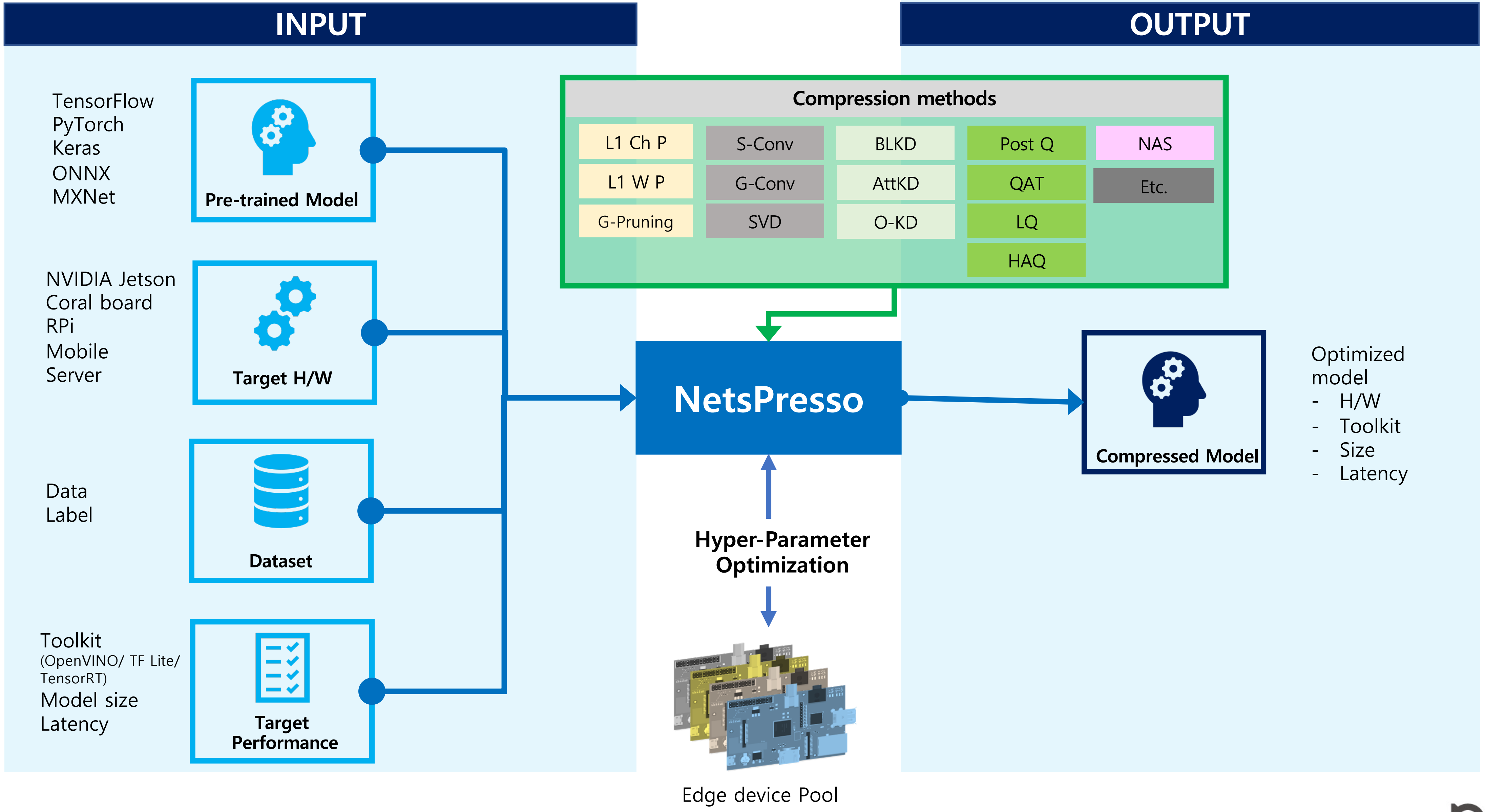
NetsPresso (Automatic Model Compression Platform)

- **Problem Solving**
  - Automatic compression without manpower
  - Combination of multiple compression methods
  - Fitted metric for practical usage

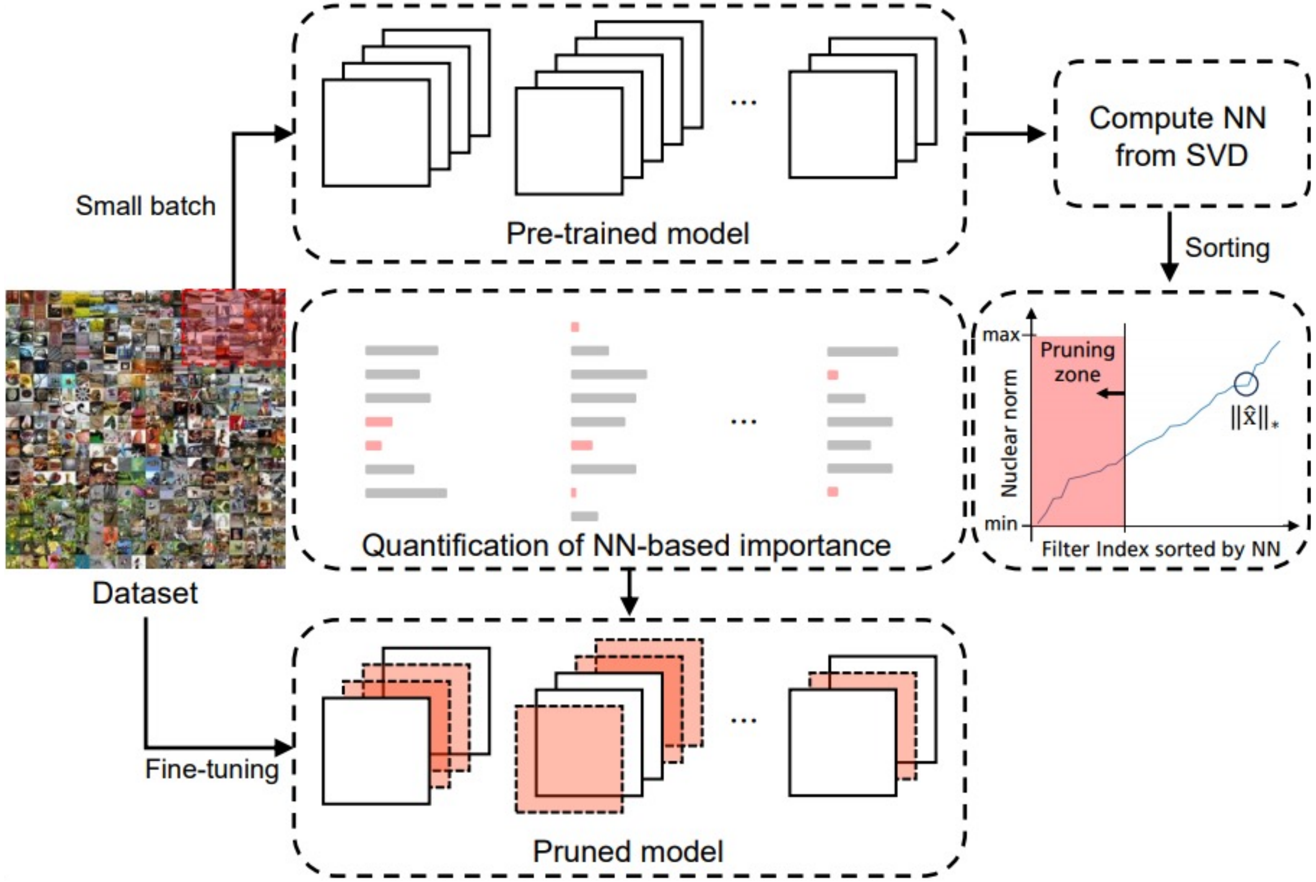
Nota's Automatic AI Model Compression Platform : NetsPresso

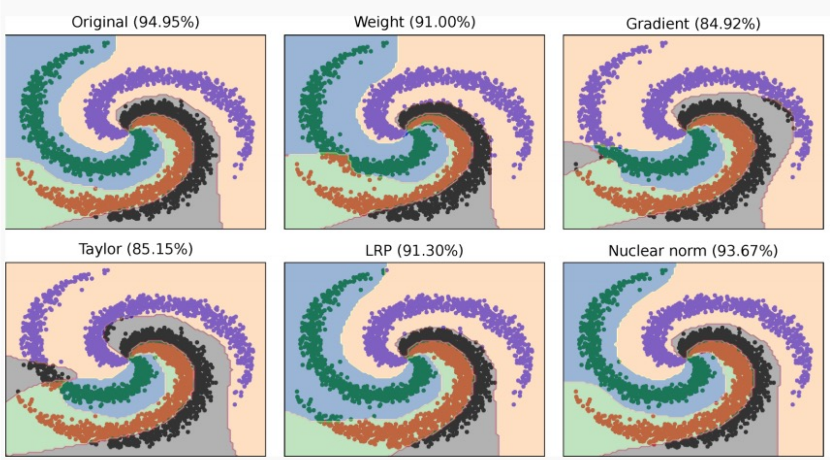
- Optimum compression platform for :
  - ✓ Target task
  - ✓ Target dataset
  - ✓ Target device
  - ✓ Target accuracy / latency / model size

Structure of NetsPresso



Toward Compact Deep Neural Networks via Energy-Aware Pruning (ICCV 2021, To be submitted)





Criterion	ResNet-50 Top-1 Acc (%) Pruned	Top-1 Acc (%) Gap	Top-5 Acc (%) Pruned	Top-5 Acc (%) Gap	FLOPs Reduc. (%)	Params Reduc.(%)
He *et al.* [11]	72.3	-3.85	90.8	-1.4	2.73B (33.25)	N/A
ThiNet-50 [33]	72.04	-0.84	90.67	-0.47	N/A (36.8)	N/A (33.72)
SSS 26 [15]	71.82	-4.33	90.79	-2.08	2.33B (43.0)	15.60M (38.8)
SSS 32 [15]	74.18	-1.97	91.91	-0.96	2.82B (31.0)	18.60M (27.0)
GAL-0.5 [29]	71.95	-4.2	90.94	-1.93	2.33B (43.0)	21.20M (16.8)
GAL-0.5-joint [29]	71.8	-4.35	90.82	-2.05	1.84B (55.0)	19.31M (24.2)
GAL-1 [29]	69.88	-6.27	89.75	-3.12	1.58B (61.3)	14.67M (42.4)
GAL-1-joint [29]	69.31	-6.84	89.12	-3.75	1.11B (72.8)	10.21M (59.9)
GDP-0.5 [28]	69.58	-6.57	90.14	-2.73	1.57B (61.6)	N/A
GDP-0.6 [28]	71.19	-4.96	90.71	-2.16	1.88B (54.0)	N/A
HRank [27]	74.98	-1.17	92.33	-0.54	2.30B (43.7)	16.15M (36.6)
HRank [27]	71.98	-4.17	91.01	-1.86	1.55B (62.1)	13.77M (46.0)
HRank [27]	69.1	-7.05	89.58	-3.29	0.98B (76.0)	8.27M (67.5)
SCOP [44]	75.26	-0.89	92.53	-0.34	1.85B (54.6)	12.29M (51.8)
SFP [8]	74.61	-1.54	92.06	-0.81	2.38B (41.8)	N/A
AutoPruner [32]	74.76	-1.39	92.15	-0.72	2.09B (48.7)	N/A
FPGM [9]	75.59	-0.56	92.27	-0.6	2.55B (37.5)	14.74 (42.2)
Taylor [34]	74.5	-1.68	N/A	N/A	N/A (44.5)	N/A (44.9)
RRBP [51]	73	-3.1	91	-1.9	N/A	N/A (54.5)
Propose method	**75.25**	**-0.89**	**92.49**	**-0.37**	**1.52B (62.8)**	**11.05M (56.7)**
	72.28	**-3.87**	**90.934**	**-1.936**	**0.95B (76.7)**	**8.02M (68.6)**

Pruning weight with nuclear norm threshold.

nota
On-device AI

Automatic Network Adaptation for Ultra-Low Uniform-Precision Quantization (IJCAI 2021, submitted)

Algorithm 1: Neural Channel Expansion

Input:
Split the training set into two dis-joint sets: D_{weight} and D_{arch} $(n(D_{weight}) = n(D_{arch}))$
Search Parameter: $\{\alpha_1^l, \alpha_2^l, .., \alpha_n^l\} \in A^l$, $\{A^1, A^2, .., A^L\} \subset \mathbb{A}$, L =number of layer
Expand Threshold: T

For Warm-up Epoch **do**
  Sample batch data D_w from D_{weight} and network from $\mathbb{A} \sim U(0,1)$
  Calculate $Loss_{weight}$ on D_w to update network weights
End for
For Search Epoch **do**
  Sample batch data D_w from D_{weight} and network from $Softmax(\mathbb{A})$
  Calculate $Loss_{weight}$ on D_w to update network weights
  Sample batch data D_a from D_{arch} and network from $Softmax(\mathbb{A})$
  Calculate $Loss_{arch}$ on D_a to update $\mathbb{A}$
  For layer **do**
    $j \leftarrow \#A^l$
    If $Softmax(\alpha_j^l; \{\alpha_k^l\}_{k \in j}) \geq T$ **do**
      Expand search space(α_{j+1}^l)
      $\alpha_{j+1}^l \leftarrow \alpha_j^l$ # copy search parameter
    End if
  End for
End for
Derive the searched network from $\mathbb{A}$
Randomly initialize the searched network and optimize it on the training set

	Accuracy	**HW Compatibility**
Uniform Precision Quantization	Low	High
Mixed Precision Quantization	High	Low
Proposed Method	High	High

Network	Method	Top-1 Acc	Top-5 Acc	FLOPs	PARAM
ResNet18	*Full precision*	*70.56%*	*89.88%*	1.814G	11.69M
	w/o NCE(Ours)	64.08%	86.47%	1.814G	11.69M
	w/ NCE(Ours)	**66.17%**	**86.75%**	**1.747G**	**12.57M**
	LSQ	67.6%	87.6%		
	QIL	65.7%	-		
	LQ-Nets	64.9%	85.9%	1.814G	11.69M
	PACT	64.4%	85.6%		
	EdMIPS	65.9%	86.5%		
ResNet50	*Full precision*	*76.82%*	*93.33%*	4.089G	25.56M
	w/o NCE(Ours)	72.36%	90.81%	4.089G	25.56M
	w/ NCE(Ours)	**74.03%**	**91.63%**	**3.932G**	**17.66M**
	LSQ	73.7%	91.5%		
	LQ-Nets	71.5%	90.3%	4.089G	25.56M
	PACT	72.2%	90.5%		
	EdMIPS	72.1%	90.6%		

Uniform precision quantization with channel expansion

nota
On-device AI

Performance

	EfficientNetB0	ShuffleNetV2		3B+ / ARM Cortex-A53 MBv3 + SSDlite
Accuracy	84.79% $\xrightarrow{+2.71}$ 87.5%	85.51% $\xrightarrow{+0.26}$ 85.77%	mAP	15% $\xrightarrow{+0}$ 15%
Number of Parameters	4.06M $\xrightarrow{x81\%}$ 3.31M	4.03M $\xrightarrow{x52\%}$ 2.09M	FLOPs	330M $\xrightarrow{x65\%}$ 215M
Inference Time	54ms $\xrightarrow{x84\%}$ 45ms	112ms $\xrightarrow{x86\%}$ 97ms	FPS	3.3 $\xrightarrow{x151\%}$ 5

- Classification tasks trained on cifar10 & ran on Intel CPU Xeon E5
- Detection task trained on COCO and ran on Rpi3B+ (1core 1thread)

Performance

AWS / Detection(Resnet34)

aws	Original		netspresso	Original		netspresso
Instance	M60	Same	M60	V100	Lower	T4
Image	5,000	Same	5,000	5,000	Same	5,000
Speed	2,300sec	59% faster	950sec	615sec	15% faster	525sec
Accuracy	26.2%	- 3.8%	25.2%	26.2%	- 3.8%	25.2%
Fee (Month)	$695	60% save	$287	$3,150	85% save	$481

ITS solution with Cameras on Nvidia Jetson Xavier

- Need to reduce traffics during rush hours and real-time traffic controls for emergency vehicles.
- 1st commercialized case in KR for on-device ITS solution. (Pyeongtaek city, Gyeonggi-do)

Demo video: CONFIDENTIAL

Demo video: CONFIDENTIAL

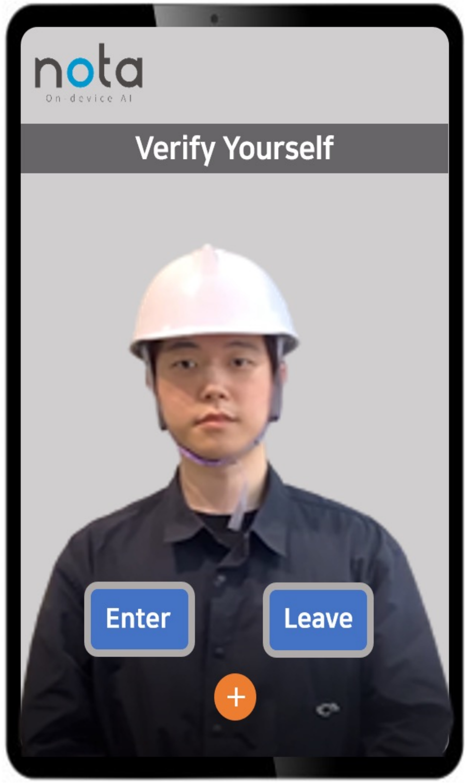
Facial Recognition with Cameras on Nvidia Jetson Nano

Project Overview

- Target to the largest construction site in Asia.
- **1/10** price of existing authentication solution
- Entry-exit tracking system in restricted areas
- Tracks the duration time of workers in restricted areas
- Manages more than **10,000 workers with an edge device**.

Highlights

- **Network-free**: Nota's AI works independently on edge devices without any network connectivity, making it a portable solution.
- **No environmental dependencies**: Our SW can be operated in diverse conditions. (low light, etc.)
- **Detection of accessory presence**: It shows the same accuracy with helmet and face mask on (21. 2Q).
- [Optional] Intermediary server: Using an intermediary server, customers can send specific data to their server in real-time and can identify users across multiple edge devices.



Inventory Management Solution with Cameras on Nvidia Jetson Xavier

- Need to reduce resources checking inventory levels in a large market.
- Collaboration with 2nd biggest retailer in KR.

[On-device inventory management (%)]

[On-device inventory management (class)]

nota
On-device AI

nota

THANK YOU FOR YOUR KIND ATTENTION

Nota Incorporated, which has a philosophy of using AI/ML to make the world more convenient, started from Korea Advanced Institute of Science and Technology(KAIST)

www.nota.ai

Nota Incorporated, Room 3104, Bldg.N28, 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
T +82 2 555 8659 l **E** contact@nota.ai l **W** http://www.nota.ai

Empowering Product Creators to Harness Edge AI and Vision

The Edge AI and Vision Alliance (www.edge-ai-vision.com) is a partnership of >100 leading edge AI and vision technology and services suppliers, and solutions providers

Mission: To inspire and empower engineers to design products that perceive and understand.

The Alliance provides low-cost, high-quality technical educational resources for product developers

Register for updates at www.edge-ai-vision.com

The Alliance enables edge AI and vision technology providers to grow their businesses through leads, partnerships, and insights

For membership, email us: membership@edge-ai-vision.com

Join us at the Embedded Vision Summit
May 25-28, 2021—Online

The only industry event focused on practical techniques and technologies for system and application creators

- *"Awesome! I was very inspired!"*
- *"Fantastic. Learned a lot and met great people."*
- *"Wonderful speakers and informative exhibits!"*

Embedded Vision Summit 2021 highlights:

- **Inspiring keynotes** by leading innovators
- High-quality, practical **technical, business and product talks**
- Exciting **demos**, **tutorials** and **expert bars** of the latest applications and technologies

Visit www.EmbeddedVisionSummit.com to learn more and register (use promo code EARLYBIRD21 by 4/16 to receive your 15%-off Early Bird Discount!)

2021 embedded VISION summit®
VIRTUAL | MAY 25-28