cDeepArch: A Compact Deep Neural Network Architecture for Mobile Sensing

Kang Yang[1], Xiaoqing Gong[1], Yang Liu[2], Zhenjiang Li[2],

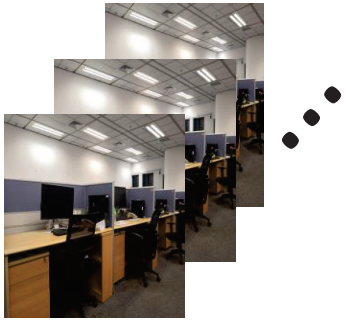
Tianzhang Xing[1], Xiaojiang Chen[1], Dingyi Fang[1]

[1]Northwest University, China

[2]City University of Hong Kong

Motivation

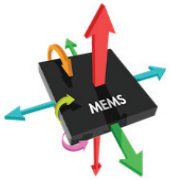
Camera



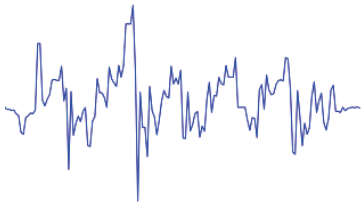
Gyro.

+

Learning Technology



Acc.



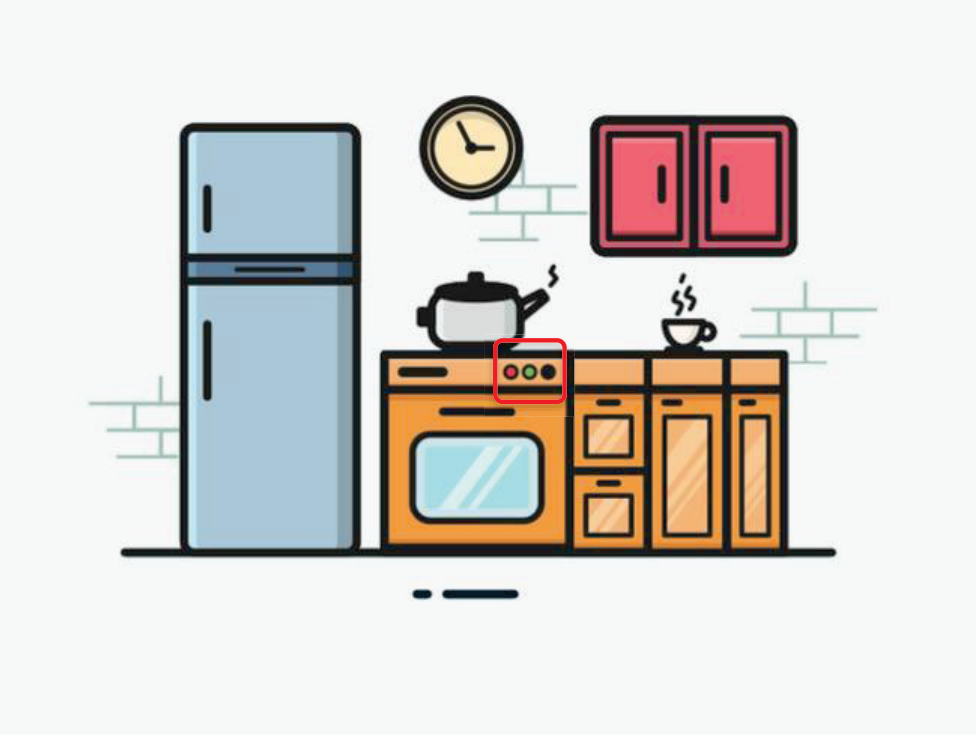
NORTHWEST UNIVERSITY • XI'AN CHINA • 1902

CityU 香港城市大學 City University of Hong Kong

Application

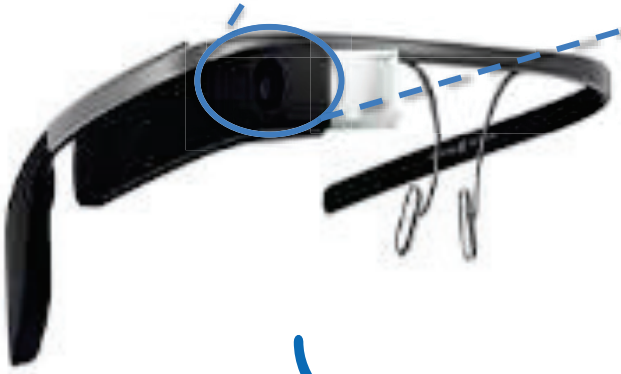
?

Cognitive decline

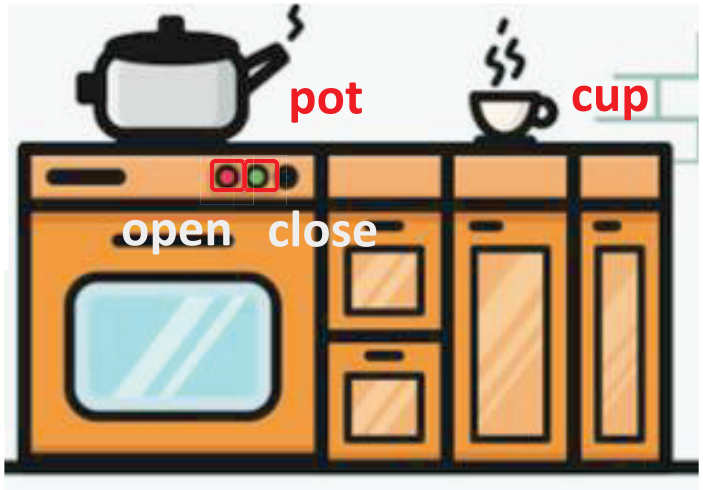


Application

First-person view



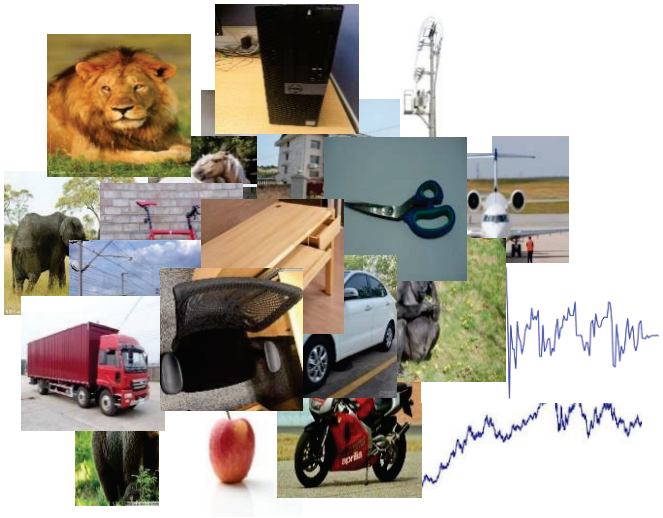
Recognizing

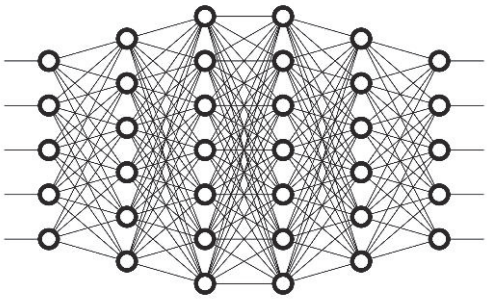


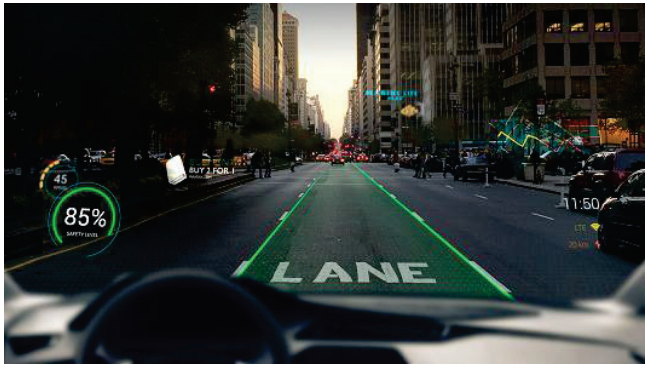
Cognitive aid system

open

Common design principle







. . .

Rich sensor data

Recognized by learning

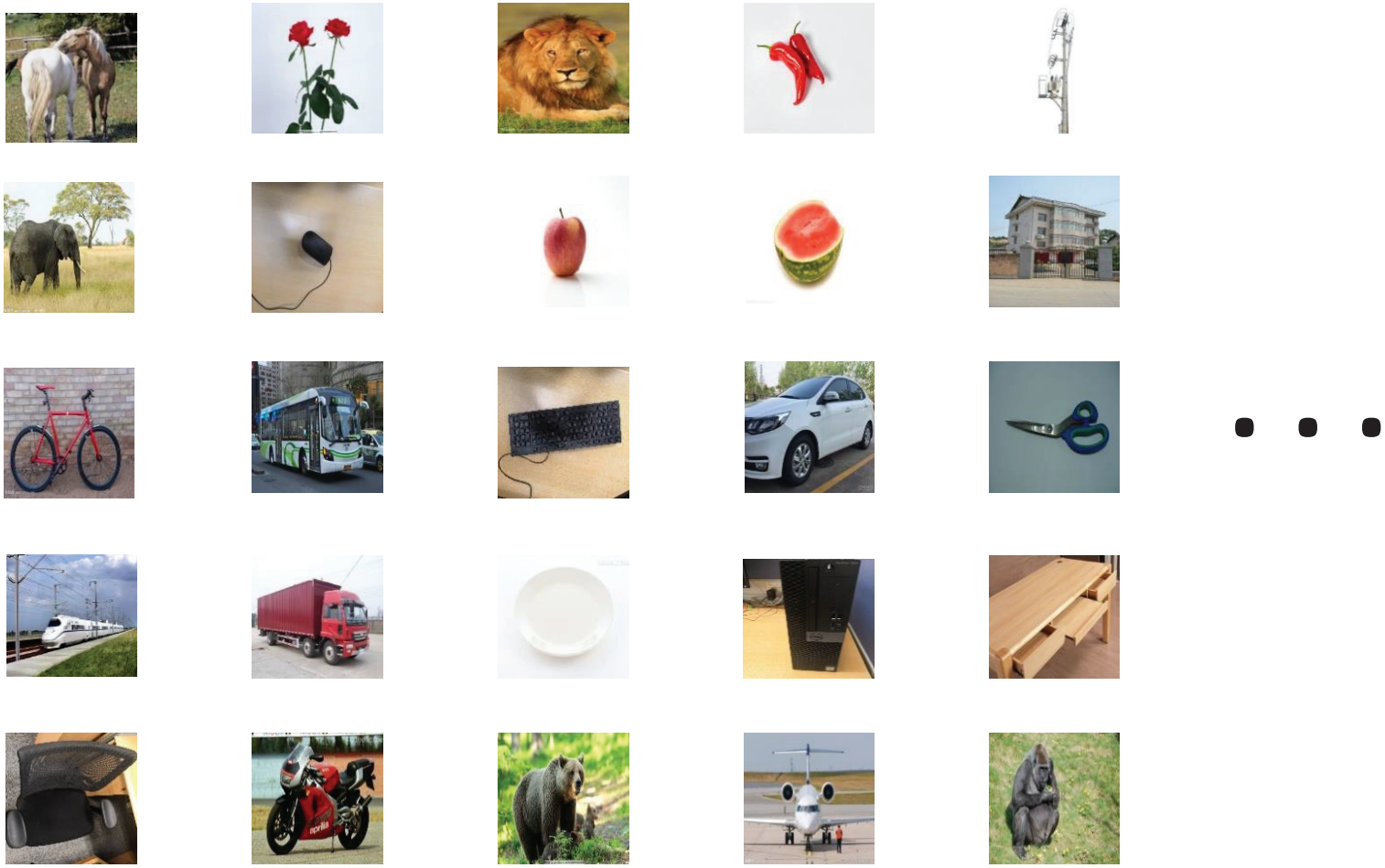
Applications

NORTHWEST UNIVERSITY · XI'AN CHINA · 1902

CityU 香港城市大學 City University of Hong Kong

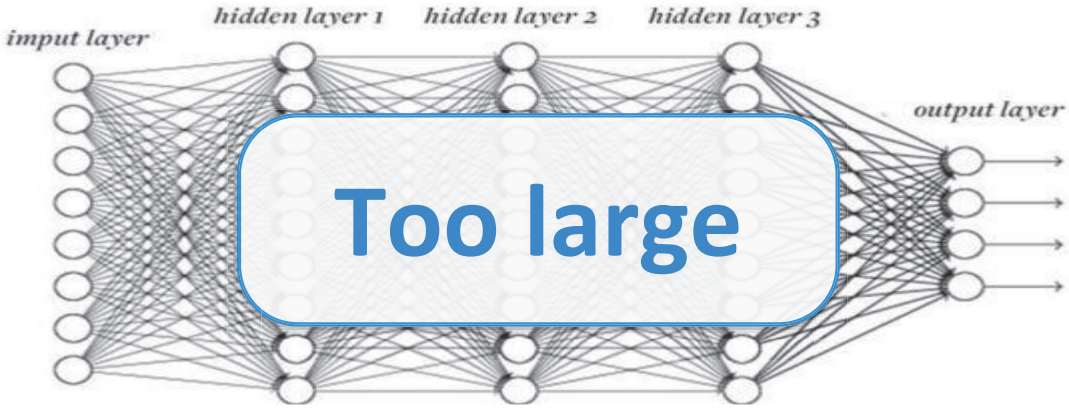
Challenges

Large targets



Challenges

- **Deep Learning**



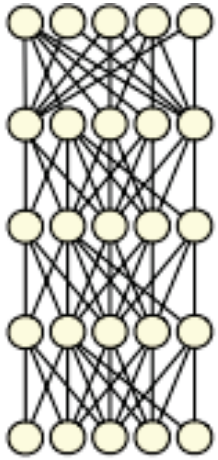
Big deep neural network

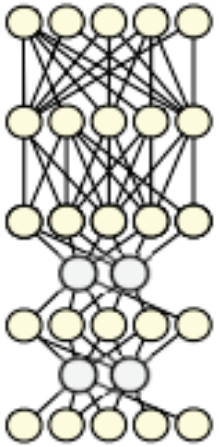
Resource-limited

Challenges

- **Deep Learning**

inaccurate



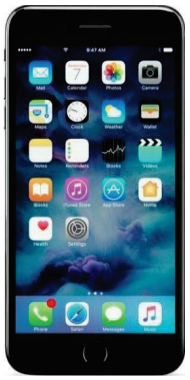


Original model

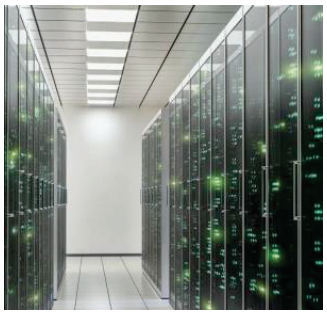
Shrunk model

No quantitative measure on available resource conditions

Any countermeasure?



0101…



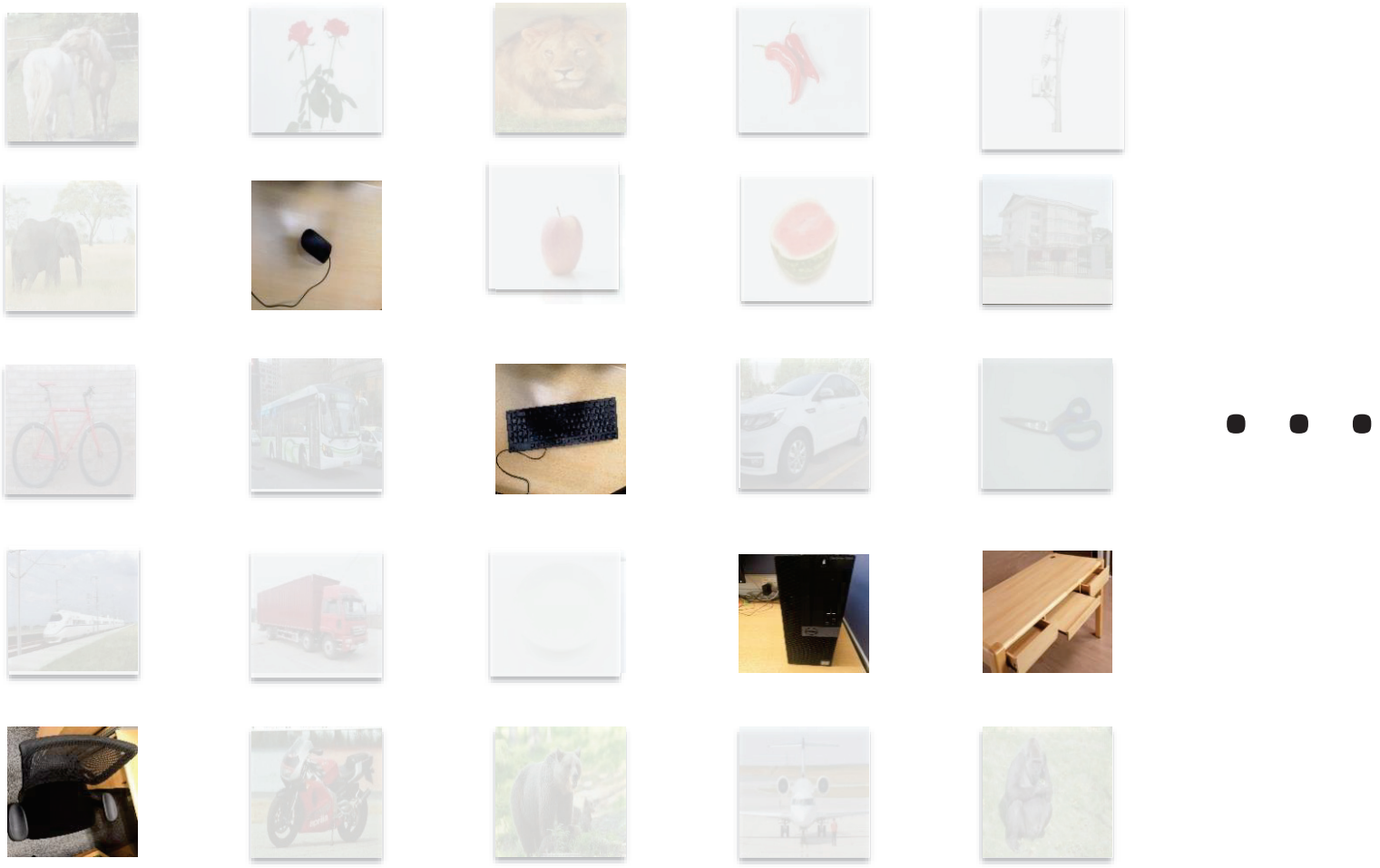
Server

Instance	Processor	vCPU	Memory(GiB)	Price($/h)
c4.large	CPU	2	3.75	0.1
c4.2xlarge	CPU	8	15	0.398
g2.2xlarge	GPU	8	15	0.65

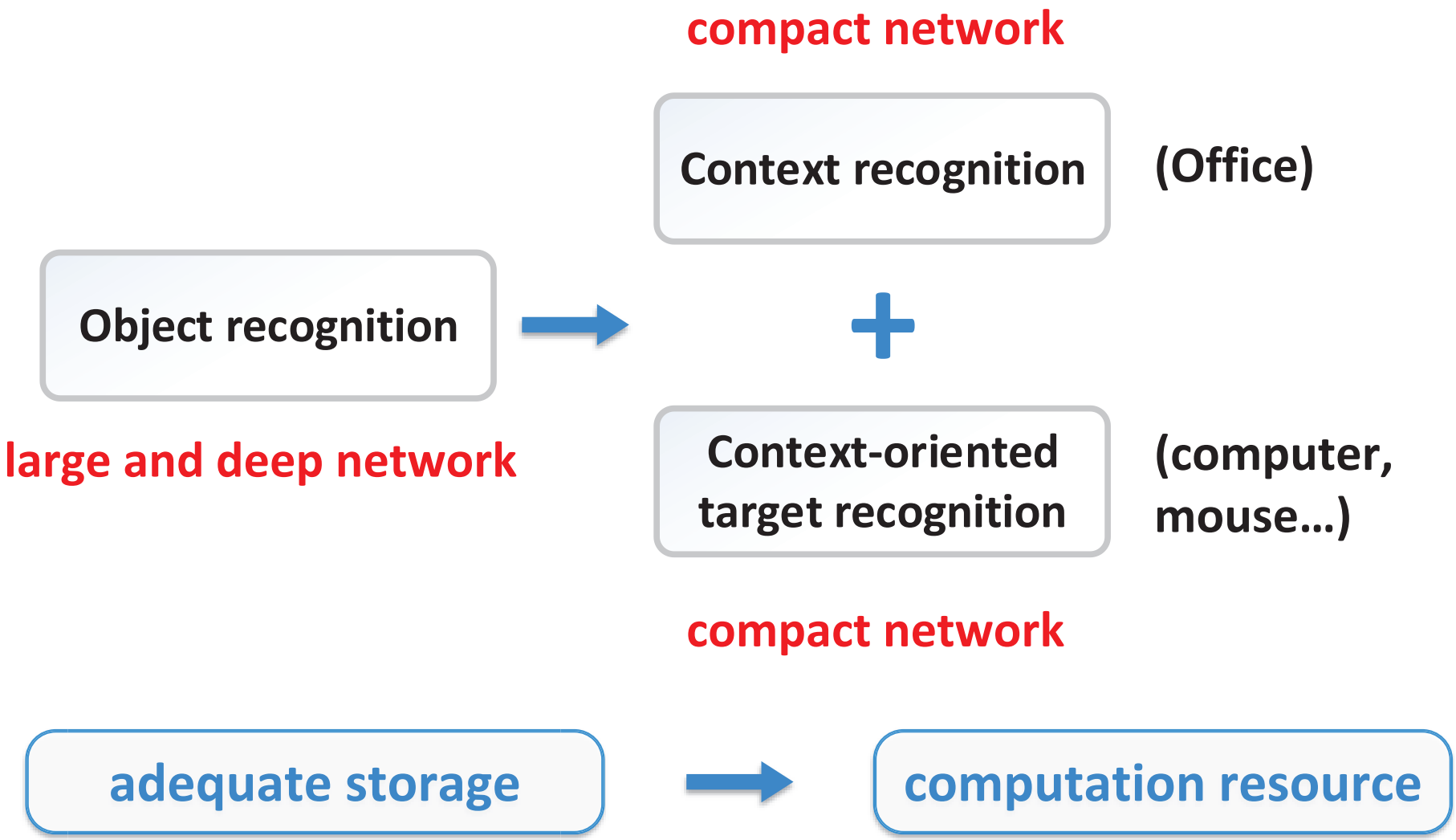
- **Long** and **uncontrollable** latency
- **High** Service **cost**
- Potential **privacy leakage**

Our solution

Context (office)



Our solution



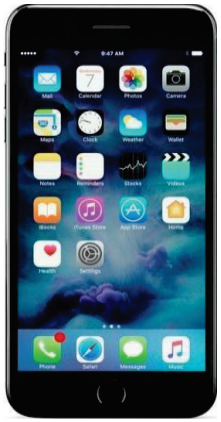
香港城市大學
City University of Hong Kong

Our solution

Context recognition

+

Context-oriented target recognition



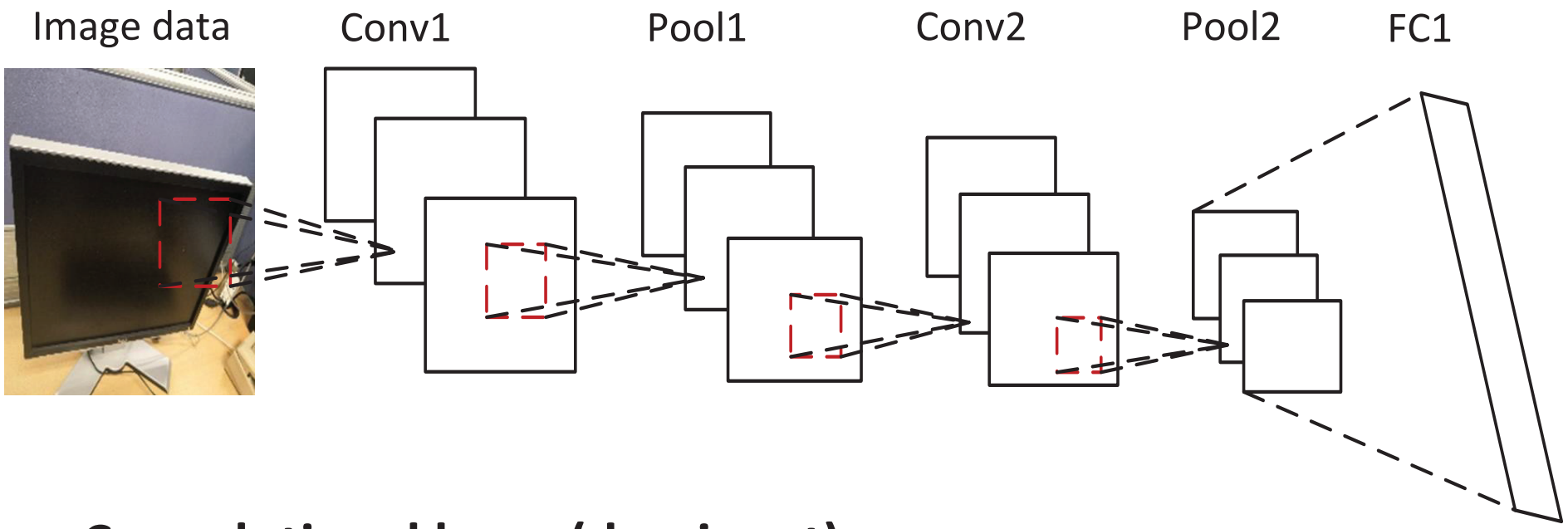
energy

computation

Available resource conditions

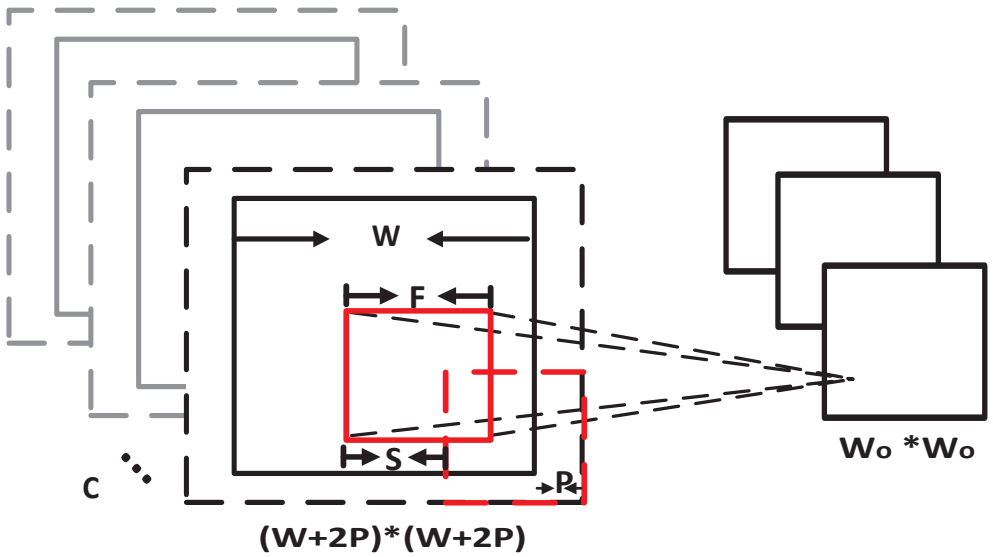
- not based on designer's experience
- Formulation facilitated configuration

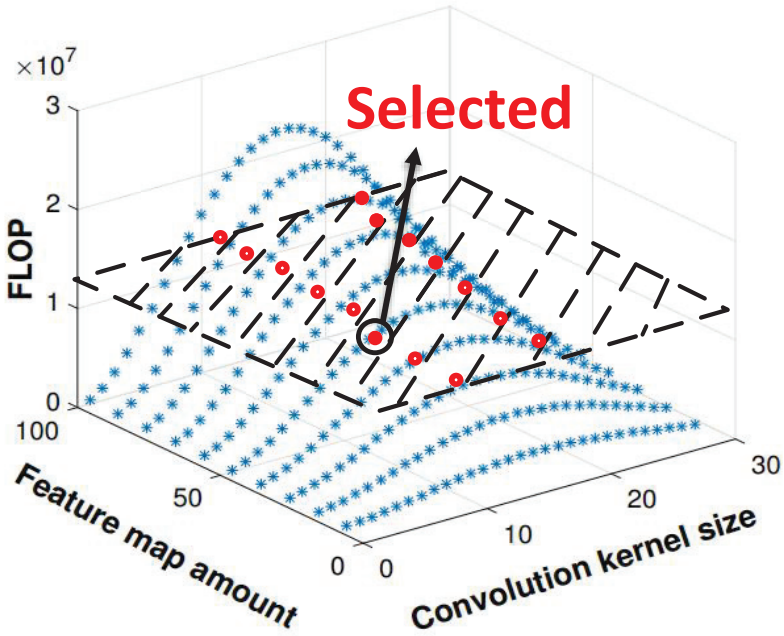
Convolutional Neural Network



- **Convolutional layer (dominant)**
- **Pooling layer**
- **Full connected layer**

Formulation facilitated configuration

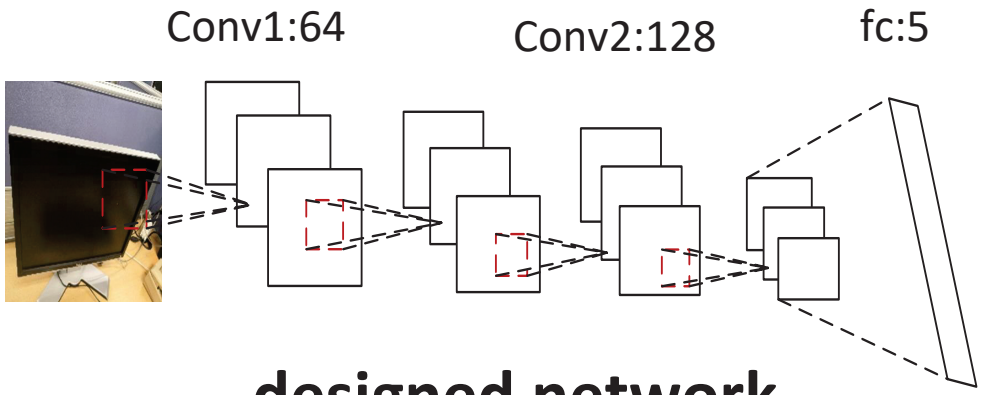




$$\mathcal{O}_{con}^{j} = W_o^j \times W_o^j \times D^j \times ((F^j)^2 \times C^j + 1),$$

$$\mathcal{O}_{con} = \sum_{j=1}^{n_{con}} \mathcal{O}_{con}^{j},$$

From computation to resource cost



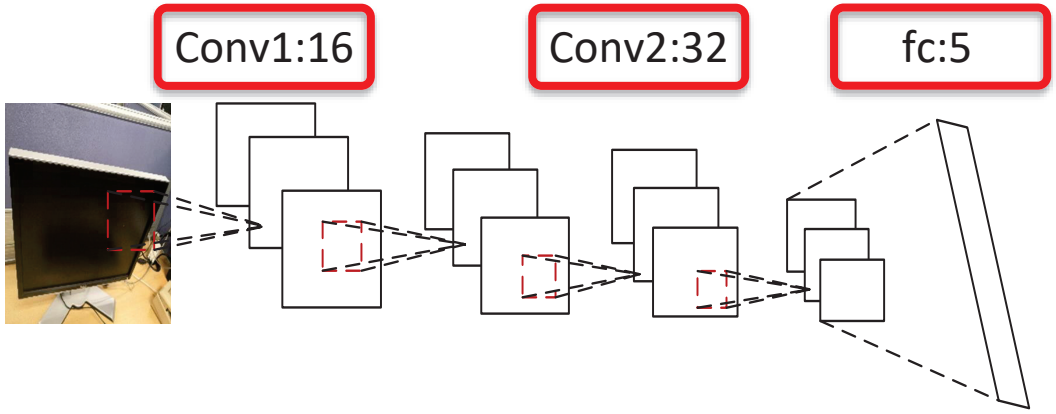
designed network

$\mathcal{O}$ computation — **Unknown**

$\mathcal{R}_i$ resource(energy)

$$\mathcal{O} = \alpha_i \times \mathcal{R}_i,$$

Derived

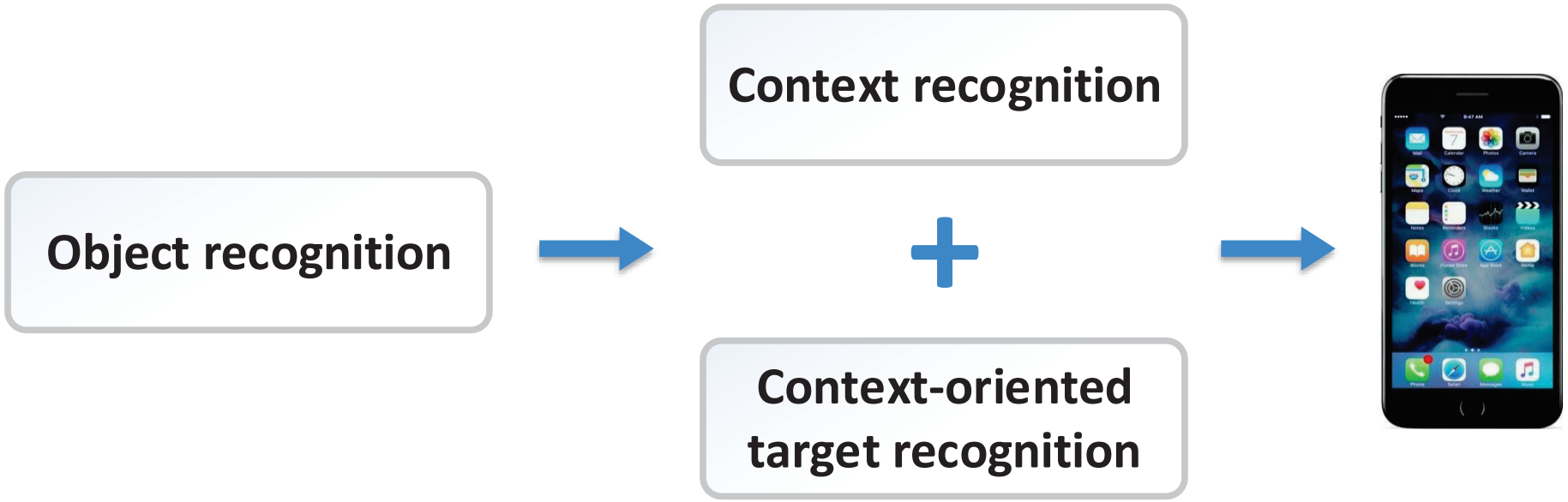


a small scale network

$\mathcal{O}$ **: computation**

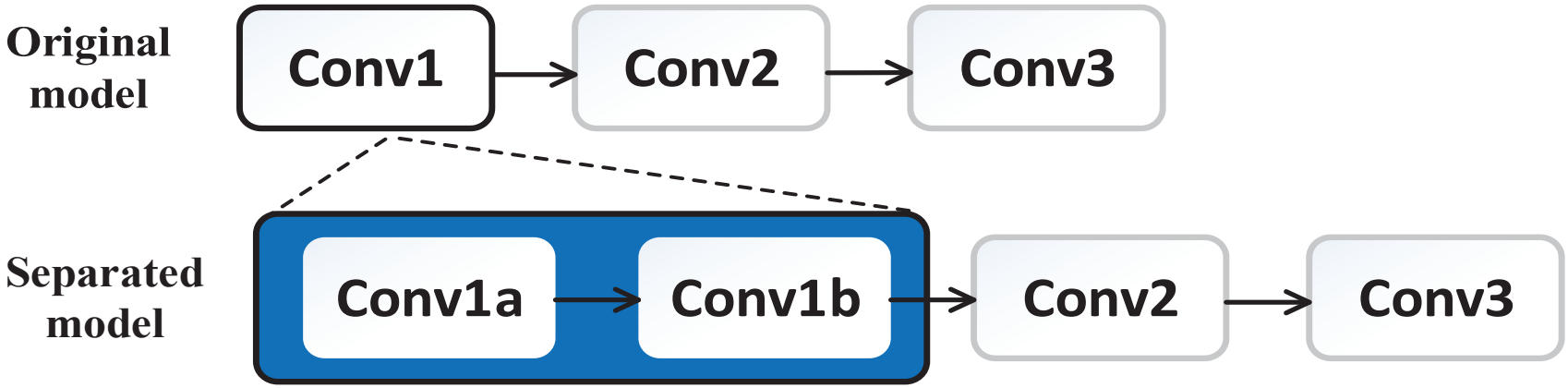
$\mathcal{R}_i$ **: actual resource consumption**

Now…



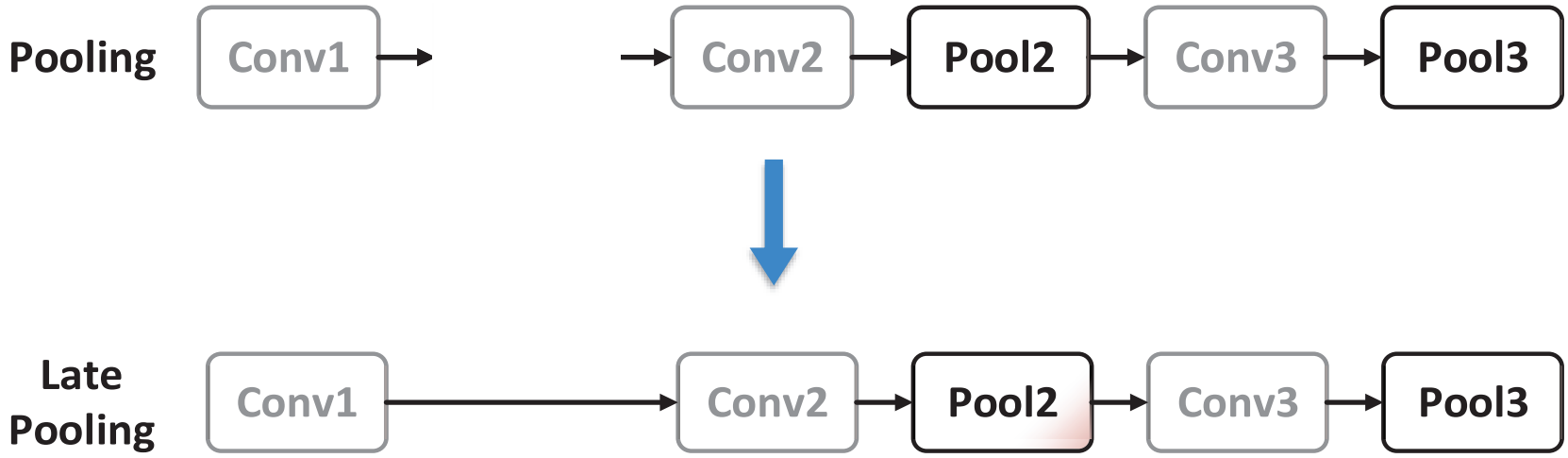
- Recognition task **decomposition**
- **Formulation facilitated** configuration
- From formulation to **estimate** the resource consumption

Enhancement: Convolutional layer

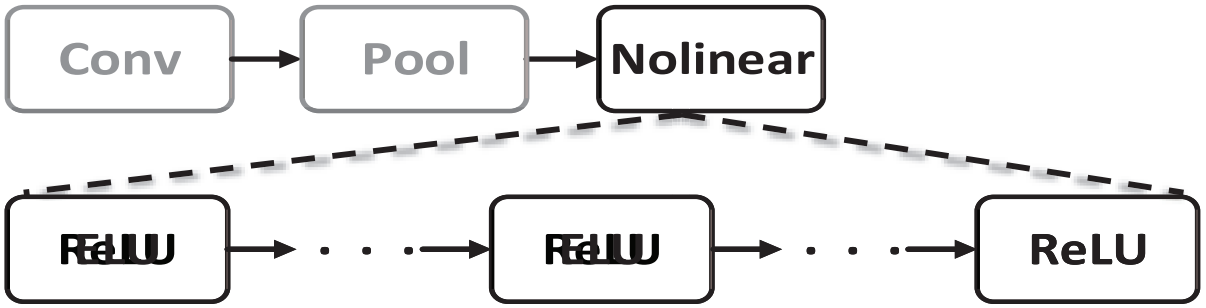


$$f \leq \sqrt{\left(F^2 - 1/C\right)/2}$$

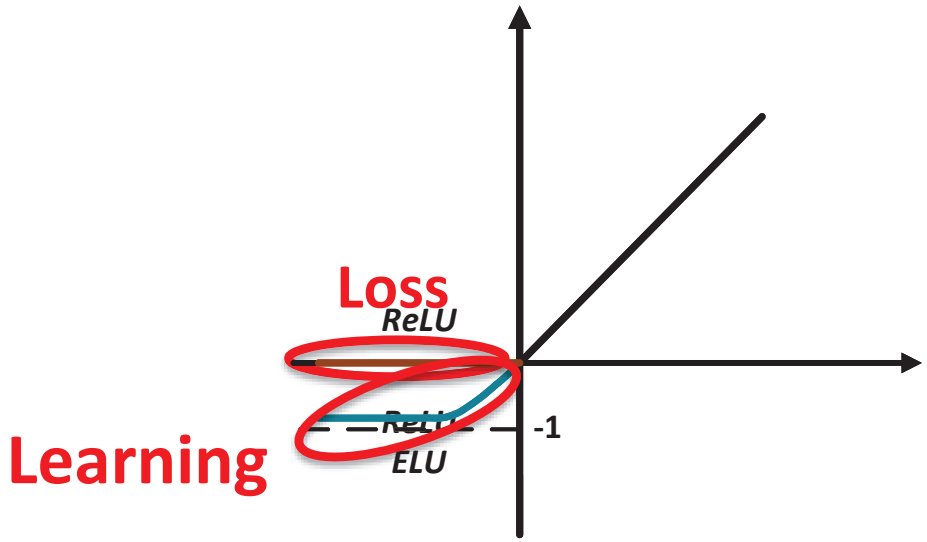
Enhancement: Pooling layer



Enhancement: Activation function



Function
ALL the Same !!!
combination



Learning

Evaluation

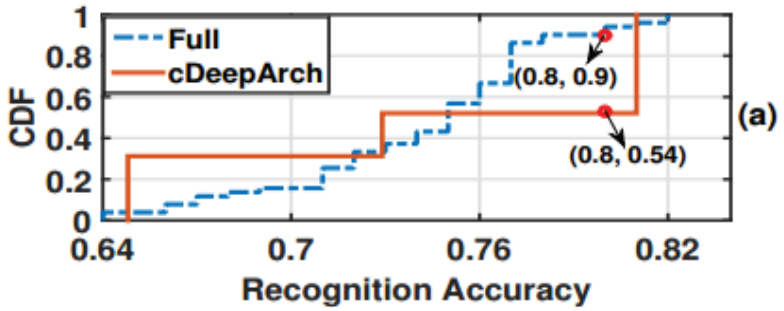
Experiments setup

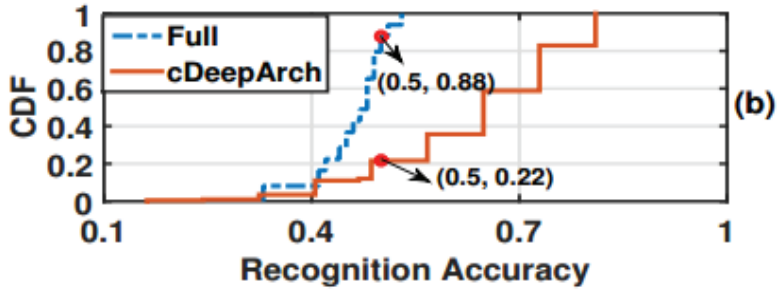
- **Dataset:**
  - **Context recognition:**
    - **MIT Place2 (related to the daily contexts)**
  - **Object recognition:**
    - **Cifar10**
    - **Cifar100 (20 classes associated contexts)**

Evaluation results

- **Overall performance**

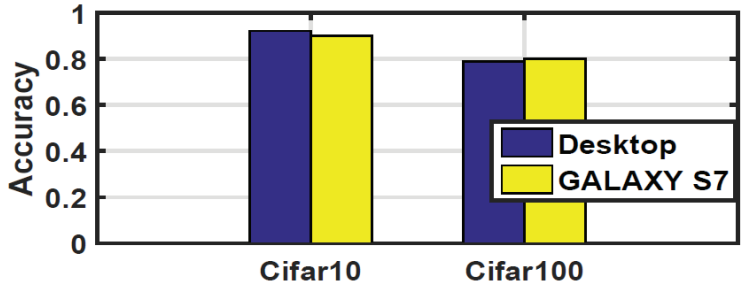
10 targets

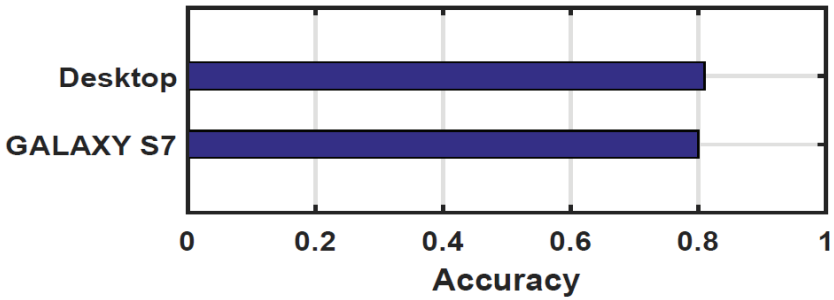




20 targets

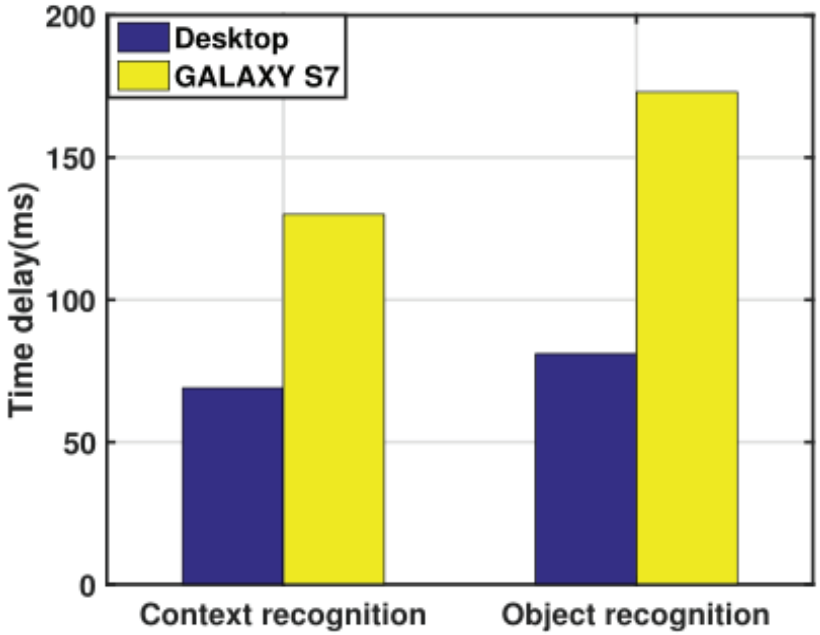
- **Recognition accuracy**





Evaluation results

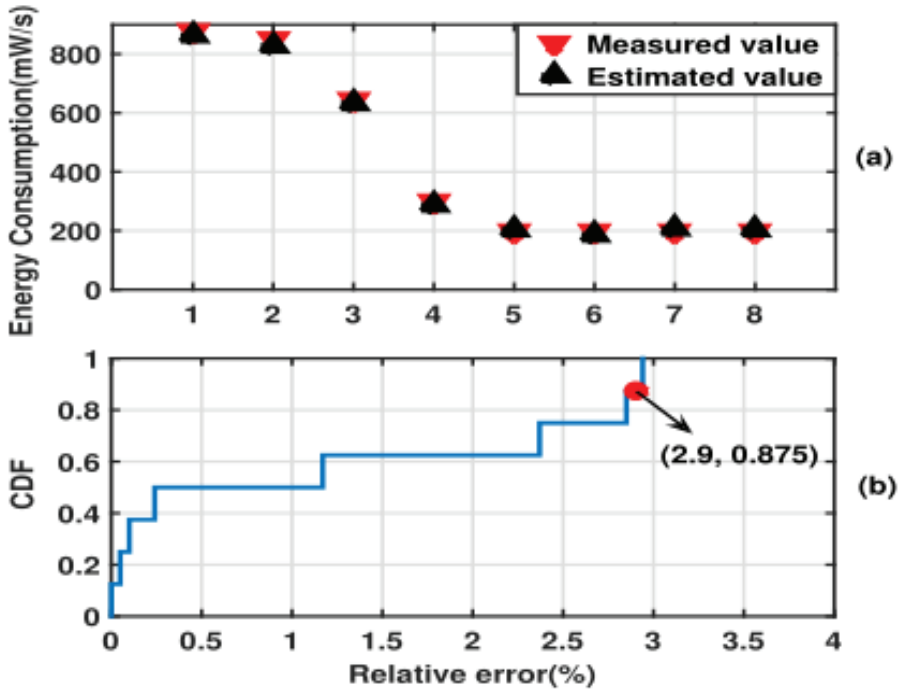
- **The time delay**



Around 150ms on Desktop

Around 303ms on GALAXY S7

- **Estimated energy values**



City University of Hong Kong

Conclusion 1, 2, 3

1. **Large targets → Decompose recognition task**
2. **Systematic way to configure network → Execution overhead formulation facilitated configuration**
3. **Enhancement techniques**

Excellent recognition performance

Lightweight

Q&A

cDeepArch: A Compact Deep Neural Network Architecture for Mobile Sensing

Kang Yang[1], Xiaoqing Gong[1], Yang Liu[2], Zhenjiang Li[2],

Tianzhang Xing[1], Xiaojiang Chen[1], Dingyi Fang[1]

[1]Northwest University, China

[2]City University of Hong Kong