cDeepArch: A Compact Deep Neural Network Architecture for Mobile Sensing

Kang Yang\textsuperscript{1}, Xiaoqing Gong\textsuperscript{1}, Yang Liu\textsuperscript{2}, Zhenjiang Li\textsuperscript{2}, Tianzhang Xing\textsuperscript{1}, Xiaojiang Chen\textsuperscript{1}, Dingyi Fang\textsuperscript{1}

\textsuperscript{1}Northwest University, China
\textsuperscript{2}City University of Hong Kong
Motivation
Application

Cognitive decline
Application

First-person view

Recognizing pot, cup, open, close

Cognitive aid system

open
Common design principle

Rich sensor data → Recognized by learning → Applications
Challenges

Large targets
Challenges

• Deep Learning

Big deep neural network

Too large

Resource-limited
Challenges

- Deep Learning

Original model  \rightarrow  Shrunken model  \rightarrow  No quantitative measure on available resource conditions
Any countermeasure?

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

### Instance Comparison

<table>
<thead>
<tr>
<th>Instance</th>
<th>Processor</th>
<th>vCPU</th>
<th>Memory(GiB)</th>
<th>Price($/h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>c4.large</td>
<td>CPU</td>
<td>2</td>
<td>3.75</td>
<td>0.1</td>
</tr>
<tr>
<td>c4.2xlarge</td>
<td>CPU</td>
<td>8</td>
<td>15</td>
<td>0.398</td>
</tr>
<tr>
<td>g2.2xlarge</td>
<td>GPU</td>
<td>8</td>
<td>15</td>
<td>0.65</td>
</tr>
</tbody>
</table>
Our solution
Our solution

Object recognition

compact network

Context recognition

(Office)

large and deep network +

Context-oriented target recognition

(computer, mouse...)

compact network

adequate storage

computation resource

adequate storage
Our solution

- **not** based on designer’s experience
- **Formulation facilitated** configuration

Context recognition + Context-oriented target recognition

Available resource conditions

Energy computation
Convolutional Neural Network

- Convolutional layer (dominant)
- Pooling layer
- Full connected layer
Formulation facilitated configuration

\[
O_{\text{con}}^j = W_o^j \times W_o^j \times D^j \times ((F^j)^2 \times C^j + 1),
\]

\[
O_{\text{con}} = \sum_{j=1}^{n_{\text{con}}} O_{\text{con}}^j
\]
From computation to resource cost

\[ O = \alpha_i \times R_i, \]

\( O \) : computation

\( R_i \) : actual resource consumption

unknown

Derived

a small scale network

designed network

Conv1:64 Conv2:128 fc:5

Conv1:16 Conv2:32 fc:5
Now...

- Recognition task **decomposition**
- **Formulation facilitated** configuration
- From formulation to **estimate** the resource consumption
Enhancement: Convolutional layer

Original model

Separated model

\[ f \leq \sqrt{(F^2 - 1/C)/2} \]
Enhancement: Pooling layer

Pooling:

Conv1 → Conv2 → Pool2 → Conv3 → Pool3

Late Pooling:

Conv1 → Conv2 → Pool2 → Conv3 → Pool3
Enhancement: Activation function

Conv → Pool → Nonlinear

ReLU → ReLU → ... → ReLU

Loss

Learning

ALL the Same !!!
Evaluation
Experiments setup

• Dataset:
  
  o Context recognition:
    ▪ MIT Place2 (related to the daily contexts)
  
  o Object recognition:
    ▪ Cifar10
    ▪ Cifar100 (20 classes associated contexts)
Evaluation results

- **Overall performance**
  - 10 targets
    - CDF
      - Full vs. cDeepArch
      - Recognition Accuracy: (0.8, 0.9)
      - (0.8, 0.54)
    - 20 targets
    - CDF
      - Full vs. cDeepArch
      - (0.5, 0.88)
      - (0.5, 0.22)

- **Recognition accuracy**
  - Cifar10
    - Desktop: 1.0
    - GALAXY S7: 0.9
  - Cifar100
    - Desktop: 0.8
    - GALAXY S7: 0.7
Evaluation results

- The time delay
  - Around 150ms on Desktop
  - Around 303ms on GALAXY S7

- Estimated energy values

![Graph showing time delay and estimated energy values](image-url)
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 Yang¹, Xiaoqing Gong¹, Yang Liu², Zhenjiang Li², Tianzhang Xing¹, Xiaojiang Chen¹, Dingyi Fang¹

¹Northwest University, China

²City University of Hong Kong