Introduction – Introduction to Computer (計算機概論)

Winston H. Hsu (徐宏民)
National Taiwan University, Taipei

September 14, 2020

Office: R512, CSIE Building
Communication and Multimedia Lab (通訊與多媒體實驗室)
http://winstonhsu.info

 学 历:
– 纽约哥伦比亚大学 电机工程博士
– 国立台湾大学 资讯工程硕士

 研究专长:
– 机器学习
– 大规模影像视讯搜寻与辨识

 经 历:
– 国立台湾大学 资讯工程学系教授 (现任)
– 国立台湾大学 NVIDIA AI Lab主持人 (现任)
– thingnario (慧景科技) 共同创办人 (现任)
– 科学工业园区审议委员会委员 (现任)
– 协助多家大型企业成立「深度学习团队」
– Visiting Researcher, Microsoft Research Redmond (2014)
– CyberLink Corp. (訊連科技) 創始工程師, R&D Manager
– Editor Board, IEEE Multimedia Mag., AE for IEEE TMM, TCSVT
Introduction to Computer

- Before going for the lecture details for the semester, let’s see what computers can do now
- We are trained to enable such exciting applications!
Drone-Based Object Localization and Counting – First-Ever Solution

[Hsieh et al., ICCV 2017]

Solve the problem in drone-view images.

Where? How many?

Drone-Based Object Counting: Spatial + Appearance Constraints

#car counts: 75
Expressing Sentiments

Image to Poetry by Cross-Modality Understanding [Cheng et al., 2017]

- Joint work with Microsoft Research Asia; deployed live in 小冰
- Learning from the 519 poets (1920~)
- Hierarchical LSTM-like models for ensuring the intra- and inter-sentence coherence
Top 3 in 2018 IEEE Signal Processing Society Video and Image Processing (VIP) Cup by EECS Unders → Traveling to Athens, Greece
Core for AI (Machine Learning)

- A prediction function $f$ for
  - Classification (分類)
  - Recommendation (or ranking) (e.g., shopping, ad, etc.)
  - Decisions for action (go, transaction, driving, etc.)
  - Translation (voice vs. text, English vs. Chinese, image vs. caption, etc.)

$$f(Y | X; \Theta)$$

- Where to deploy?
- target
- current observations
- parameters learned (from data)

object detection
Amazing Technologies are Enabled by Tiny Bits!

Early Computers
Early Programming Tools

First Popular PCs
Early PCs

- Intel 8086 processor
- 768KB memory
- 20MB disk
- Dot-Matrix printer (9-pin)

GUI/IDE

```pascal
type pstiva = record
  next : pstiva;
  val  : longint;
end;

var
  a : array[1..100,1..100] of longint;
  d,p : array[1..100] of longint;
  n : longint;
  prim,ultim : pstiva;

procedure AddToStiva(i:longint);
begin
  if (prim = nil) then
  begin
    new(prim);
    ultim := prim;
    prim^.next := nil;
  end
  else
    if a[i,j] <> 0 then
      begin
        new(prim);
        ultim := prim;
        prim^.next := nil;
      end
  end;
```

Help: Insert a watch expression into the Watch window
More Advanced Architectures – Multi-Cores, GPUs

- Dual Next Generation Tegra
- Dual Discrete GPUs
- 12 CPU Cores
- Pascal GPU
- 8TFLOPS (FP32)
- 24DL TOPS
- 12 simultaneous LVDS camera inputs

More Advanced Software
More “Computers” around Us

DGX-1, AI Supercomputer, with 8 V100 GPUs (upgraded from P100 in 2019; the first one in Taiwan, 2016)

- 960 TFLOPS (GPU FP16)
- 40,960 CUDA Cores
- 5,120 Tensor Cores
The Downside

- “Once upon a time, every computer specialist had a gestalt understanding of how computers worked. … As modern computer technologies have become increasingly more complex, this clarity is all but lost.” Quoted from the textbook

How Does a Program Work?

```python
Python 2.7.10 (default, Feb 7 2017, 00:00:15)
[GCC 4.2.1 (Apple Inc. Build 12655) (dot 3)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> >>> >>> >>> >>> >>> >>> >>> print "hello world!"
hello world! >>> >>> >>> >>> for i in range(10):
... print "hellow world!"
... hellow world! hellow world! hellow world! hellow world! hellow world! hellow world! hellow world! hellow world! hellow world! hellow world! >>>
```
Layers of Applications and (Software/Hardware) Systems

Facebook as a three-tier system

Many others are missing here; for example, compilers, architecture, programming languages, data structure, database, etc.

Layers of Computer Cores To Be Covered

A typical Von Neumann machine

An ALU is a fundamental building block of many types of computing circuits, including the central processing unit (CPU) of computers, FPUs, and graphics processing units (GPUs).

Why Learning So Many Details?

Q: Computers are getting easier. Why I need to learn such complex details?

Answers:
- You are in the best computer science department
- You need to know how to design (both hardware/software) from the scratch
- Or Have the ability to invent the new computation paradigm
  - e.g., NOSQL, GPU-enabled data computing, quantum computer, intelligent X, etc.
Especially, Designing (Future) AI-Savvy Products – Combining both Hardware/Software

V.S.

<table>
<thead>
<tr>
<th>Operation</th>
<th>Energy [pJ]</th>
<th>Relative Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 bit int ADD</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>32 bit float ADD</td>
<td>0.9</td>
<td>9</td>
</tr>
<tr>
<td>32 bit Register File</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>32 bit int MULT</td>
<td>3.1</td>
<td>31</td>
</tr>
<tr>
<td>32 bit float MULT</td>
<td>3.7</td>
<td>37</td>
</tr>
<tr>
<td>32 bit SRAM Cache</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>32 bit DRAM Memory</td>
<td>640</td>
<td>6400</td>
</tr>
</tbody>
</table>

Energy table for 45nm CMOS process

Related Work

Neural networks are typically over-parameterized, and there is significant redundancy for deep learning models. This results in a waste of both computation and memory. There have been various proposals to remove the redundancy: Vanhoucke et al. explored a fixed-point implementation with 8-bit integer (vs 32-bit floating point) activations. Denton et al. exploited the linear structure of the neural network by finding an appropriate low-rank approximation of the parameters and keeping the accuracy within 1% of the original model. With similar accuracy loss, Gong et al. compressed deep convnets using vector quantization. These approximation and quantization techniques are orthogonal to network pruning, and they can be used together to obtain further gains.

Course Goal

- Introductory course for the (new) undergraduates in the computer science department – with strong CS-level requirements
- We will cover basic knowledge about computer, data manipulation & abstraction, computer architecture, organization, software, operating system, database, network, GitHub, machine learning, etc. (if time permitted)
- Mix with technical details and fun!
- Discovering future career path
- Expecting the students to be able to work as an (basic) intern in Summer 2021
Topics To Be Covered (Tentative)

- Data and number systems
- Boolean logic
- Boolean arithmetic (ALU)
- Sequential logic
- Machine language
- Computer architecture
- Assembler
- Operation systems
- GitHub & python programming
- networks and the Internet
- database systems
- machine learning and deep neural networks

Lecture and Homework Schedule (Tentative)

<table>
<thead>
<tr>
<th>date</th>
<th>planning</th>
<th>hw assigned</th>
<th>hw due</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/14/20</td>
<td>introduction</td>
<td>HW #1</td>
<td></td>
</tr>
<tr>
<td>9/21/20</td>
<td>data and number</td>
<td>HW #2</td>
<td></td>
</tr>
<tr>
<td>9/28/20</td>
<td>boolean logic -1</td>
<td></td>
<td>HW #1</td>
</tr>
<tr>
<td>10/5/20</td>
<td>boolean logic -2</td>
<td>HW #3</td>
<td>HW #2</td>
</tr>
<tr>
<td>10/12/20</td>
<td>sequential logic -1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/19/20</td>
<td>sequential logic -2</td>
<td>HW #4</td>
<td>HW #3</td>
</tr>
<tr>
<td>10/26/20</td>
<td>Arithmetic Logic Unit (ALU)</td>
<td>HW #5</td>
<td>HW #4</td>
</tr>
<tr>
<td>11/2/20</td>
<td>computer architecture</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/9/20</td>
<td>Midterm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/16/20</td>
<td>machine language</td>
<td>HW #6</td>
<td>HW #5</td>
</tr>
<tr>
<td>11/23/20</td>
<td>python + git</td>
<td>HW #7 python</td>
<td></td>
</tr>
<tr>
<td>11/30/20</td>
<td>TA time for python + github</td>
<td></td>
<td>HW #6</td>
</tr>
<tr>
<td>12/7/20</td>
<td>operation systems</td>
<td>HW #8</td>
<td>HW #7 python</td>
</tr>
<tr>
<td>12/14/20</td>
<td>networks + internet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/21/20</td>
<td>database (1)</td>
<td>HW#9</td>
<td>HW #8</td>
</tr>
<tr>
<td>12/28/20</td>
<td>database (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/4/21</td>
<td>machine learning + dnn</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/11/21</td>
<td>Final</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
:: Administrative Issues ::

Administrative Issues

- ~9 assignments in the lecture (every two weeks)
  - Homework (#1) available now (creating your own CV)
    - DUE: at 12pm, September 28, 2020; CEIBA submission
- TAs
  - 鍾起鳴 <r09944021@ntu.edu.tw>, R506
  - 林承緯 <r09922078@ntu.edu.tw>, R506
  - Office hours: (1) 3:30pm – 5:30pm, Wednesday,
    (2) 10am – 12pm, Thursday
- Course information
  - Course outline
    https://winstonhsu.info/2020f-comp-intro/
  - Readings, homework, slides, etc.
    https://ceiba.ntu.edu.tw/1091intro_comp/
Textbook and References

- Textbook:
  - The Elements of Computing Systems, Noam Nisan and Shimon Schocken
    - The first 6 chapters (to be used) are available online; you probably need not buy the hard copy
    - http://www1.idc.ac.il/tecs/plan.html
    - Digital copy can be easily found online

- References
    - To be used in the second half
    - 前幾屆的單班用書

Grading (Tentative)

- Assignments: 30-35%
- Midterm Exam.: 30%
- Final Exam.: 30%
- Class Participation: 5-10%

- Cheating in homeworks or exams will cause the grade as “F” automatically
Create Your Own CV (Curriculum Vitae)

- What’s CV?
- Why needing CV?
- How to fill in the perfect CV for your future career? study, job market?
- Draft one (HW #1) and keep updating for the coming years
  - Create your own webpage or GitHub
  - Be careful for your ID style – being professional
- Consult others and Google what kind of CV highlights you need?
- Tutorial (many others)
  - https://www.naaree.com/resume-writing-help/
  - https://medium.com/duomly-blockchain-online-courses/how-to-write-a-resume-for-it-professional-tutorial-with-resume-example-template-cbdc0275a16d

How to Prepare for a Good CV?

Format (Suggested)

<table>
<thead>
<tr>
<th>Left column (30%-40%)</th>
<th>Right column (60%-70%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal data</td>
<td>About me (or Objective)</td>
</tr>
<tr>
<td>Languages (optional)</td>
<td>Write 2-3 sentences about who you are, your main career path, and what you would like to do.</td>
</tr>
<tr>
<td>Natural languages</td>
<td>Main skills</td>
</tr>
<tr>
<td></td>
<td>Programming languages</td>
</tr>
<tr>
<td></td>
<td>IT projects experience</td>
</tr>
<tr>
<td></td>
<td>Non-IT projects experience</td>
</tr>
</tbody>
</table>

- Remember to proofread and edit your typos
- Choose a font size of at least 10 and avoid using too many different font to ensure the readability
- Focus more on the messages than designs
- Always keep your CV up to date