FAET630004: Computer & AI Architecture **(Due: 6/20/19)**

Homework Assignment #4 & Final Project

Instructor: Chixiao Chen *Name:* *, FudanID:*

- This HW assignment merges with final projects, counting 40 percent , please treat it carefully.
- Please prepare a slides and your code before 06/19/2019 11:59pm.
- It is encouraged to use LaTeX to edit it, the source code of the assignment is available via: https://www.overleaf.com/read/wxbbzmqscjkd
- The project includes two options, please choose one between two.
- No matter whether option you choose, it requires all your achievement from HW 1 to 3. It is highly recommended to revise your homework assignments first.
- I will arrange an on-class presentation to evaluate your works and discuss.

In the final project, assuming you are an engineer in charge of a deep learning processor in Hauwei. Please use the simplified single cycle RV32I processor in HW2 as the base line, and design a specific instruction/accelerator to speedup your processor's computing for the neural network from HW3. Pick up 5 correct inference result and perform the entire inference on your design. Answer the following questions after your design.

(a) Comparing with the design you have in a typical RV32I design, how much performance you have improved by your design ?

(b) What's your ideal utilization, and what's your real utilization in terms of MAC-Operation per second ? Discuss the gap.

(c) Compared with the ideal python computing result, is your implementation still correct? What's the score difference between your result and the python result?

(d) Hint: You are allowed to shrink/compress your network to reduce the workload, also assuming you have a big enough memory.

You can choose one of the following options to complete your project. Please submit your slides after presentation.

Option 1. Develop SIMD instructions in RV32I (baseline implementation)

Quantize your network into 8-bit, and design two 4-way SIMD instruction to implement the MAC. Treat each 32-bit register as 4 8-bit. Also, your should design a code generator to generate all the assembly code automatically. After your implementation, please to some literature research and propose one method to further improve your performance.

Option 2. Find your own method to improve the performance (innovative implementation)

Develop a more advanced method including data flows, data sparsity, or quantize it more aggressively etc., to design a more efficient processor for neural network. You have to elaborate that the method is better than option 1.