Speech Recognition

Team06

Tsang-Yung Wu Chen-Chia Chang Cynthia Y Liu

Outline

- Project Idea
- About CNN
- Workflow
- Structure of System
- Software Design
- Calculating Implement
- DEMO
- Future Work

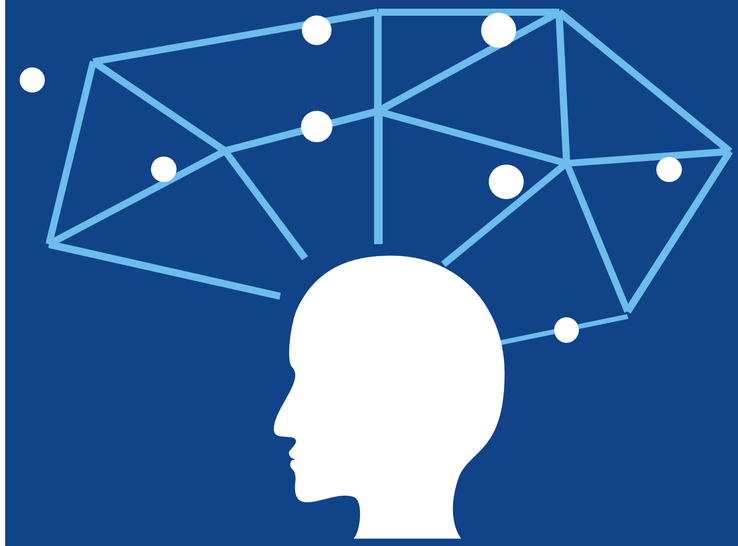
Project Idea

1

"There is only 21 hours a day, and the remaining 3 hours are for sleeping."

Speech Recognition

- Machine learning is widely used
- Often using GPU parallel operation
- Take advantage of FPGA hardware acceleration
- Implement a speech recognition system on the FPGA
- CNN

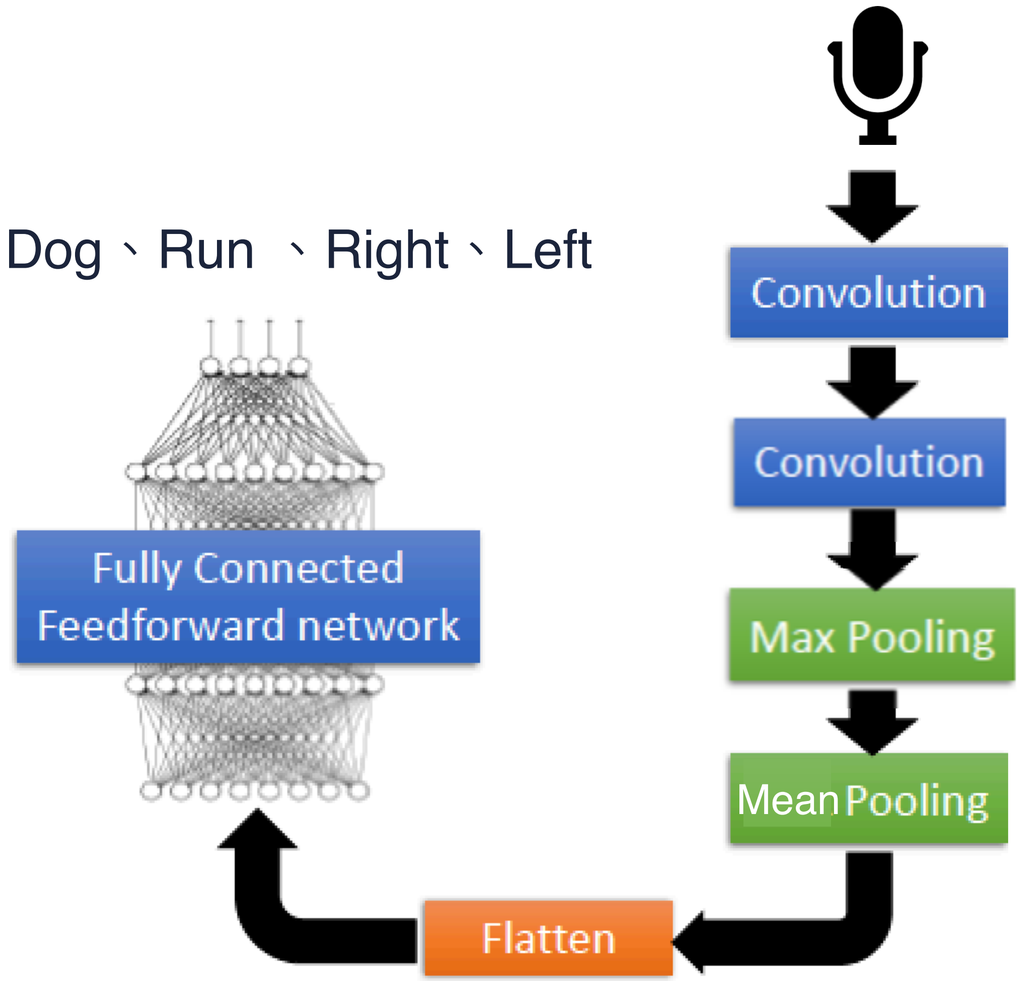


About CNN

2

"I just want to live in the lab, or else?"

About CNN



- Convolution Neural Network
- Can reduce the amount of weight compared to DNN
- We use it to implement speech recognition:

  Convolution * 2

  Max pooling

  Mean pooling

  Fully connected

Python predict

DCNN

Current model : mfcc-based with relu

precision: 8 bit, clip at 7.9375

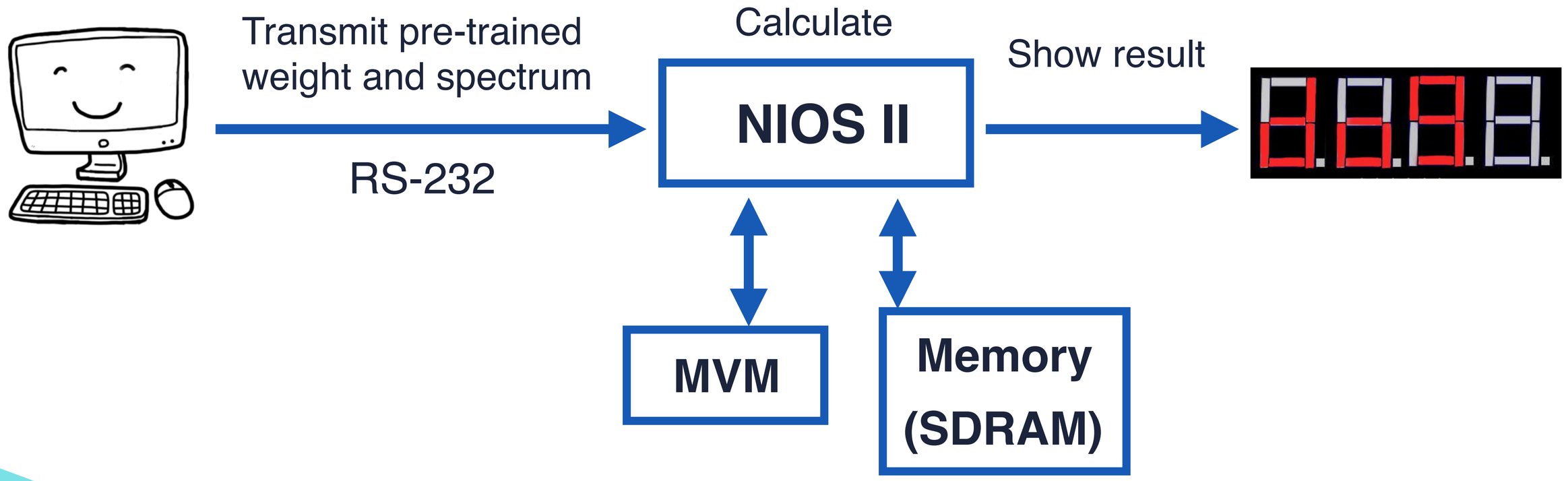
Take 200 test data as the benchmark::

	precision	time
General Version	81.5%	35.2s
Weight truncate	82%	34.5s
Clip	80.5%	34s
Clip + Weight truncate	81.5%	32s

Workflow 3

“You’re too strong."

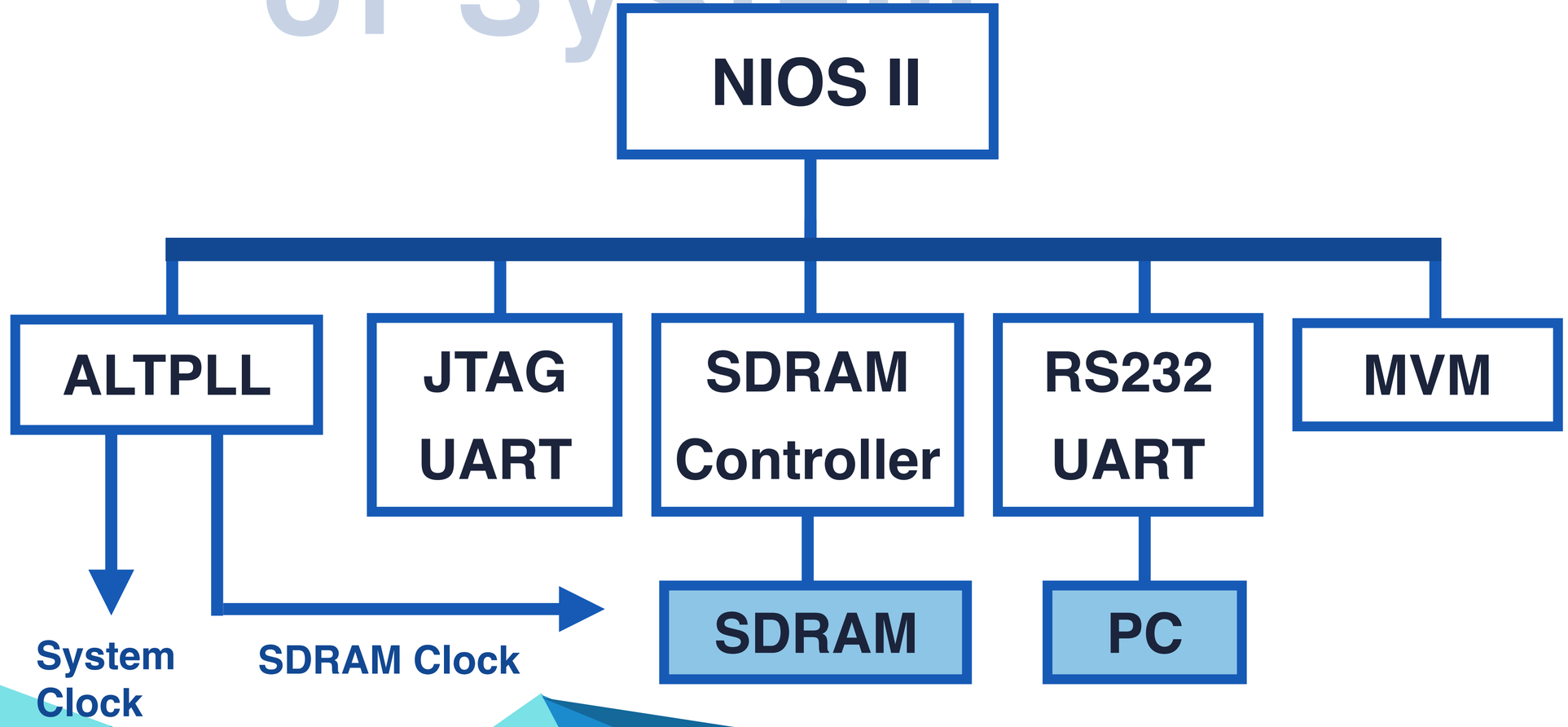
Workflow


Transmit pre-trained
weight and spectrum
RS-232
Calculate
NIOS II
Show result
MVM
Memory
(SDRAM)

Structure of System

4

Structure of System



Structure of System

- In order to use NIOS as the master, put all slaves into QSYS
- PLL synthesizes 75MHz clock, while SDRAM requires 3 ns faster
- Set the written MVM (hardware calculation) to Avalon slave and connect to NIOS.

Software Design

5

Software design

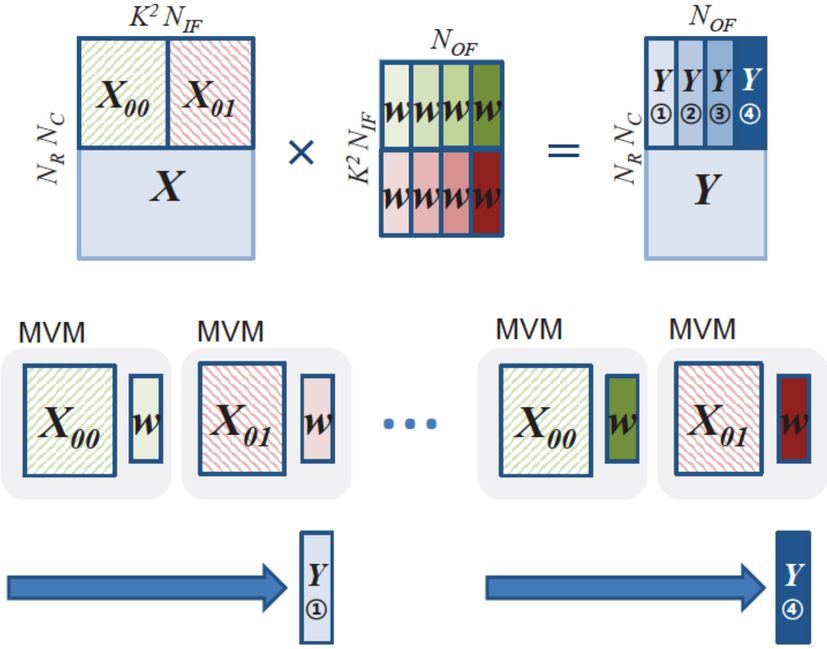
NIOS II：

- C code Implementing CNN's flow control
- Memory using SDRAM (128MB)
- Use RS232 to transmit pre-trained weight for storage in SDRAM

Software design

CNN Flow :

- Flatten matrix
- Convolution two times
  - matrix block
  - multiplication(use hardware)
  - reshape and truncate
- Max pooling
- Mean pooling
- Fully connected

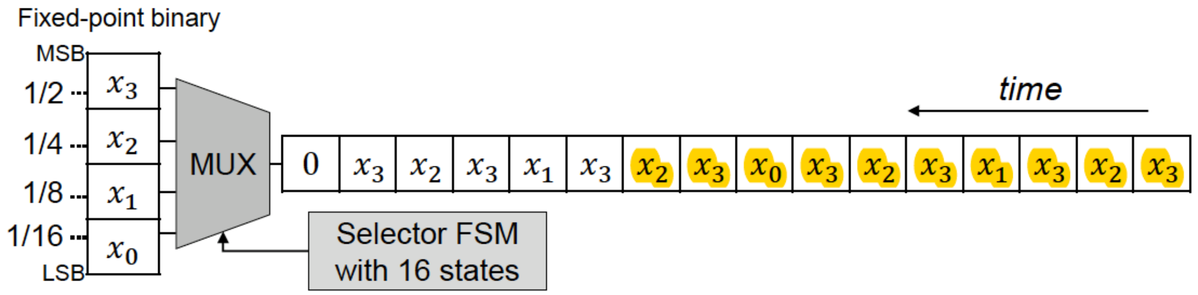


Calculating Implement

6

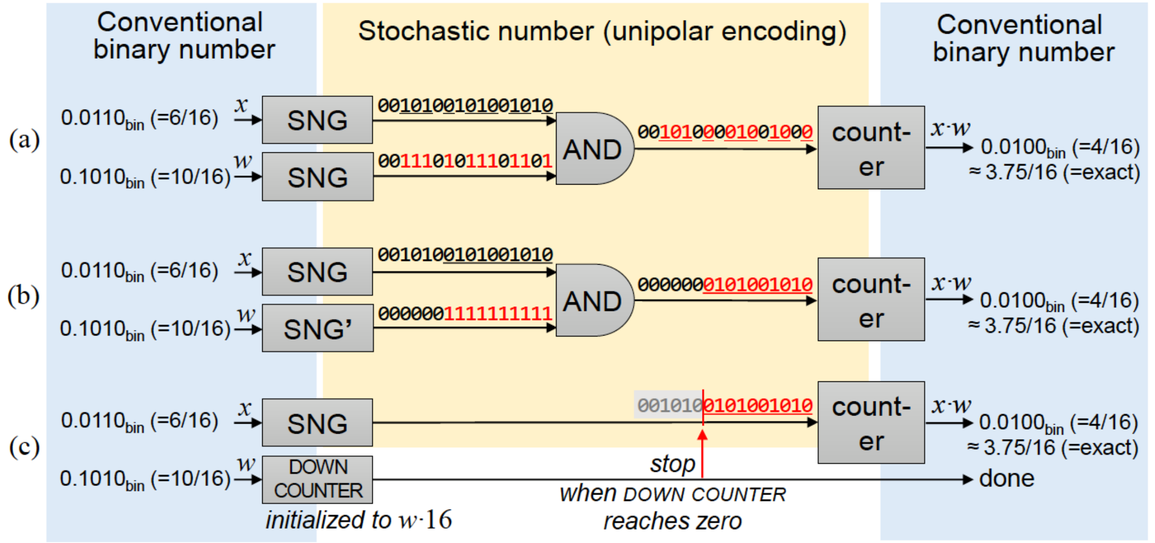
MVM Controller

- Implement fractional multiplication
- Reduce the required arithmetic unit
- Stochastic number generator



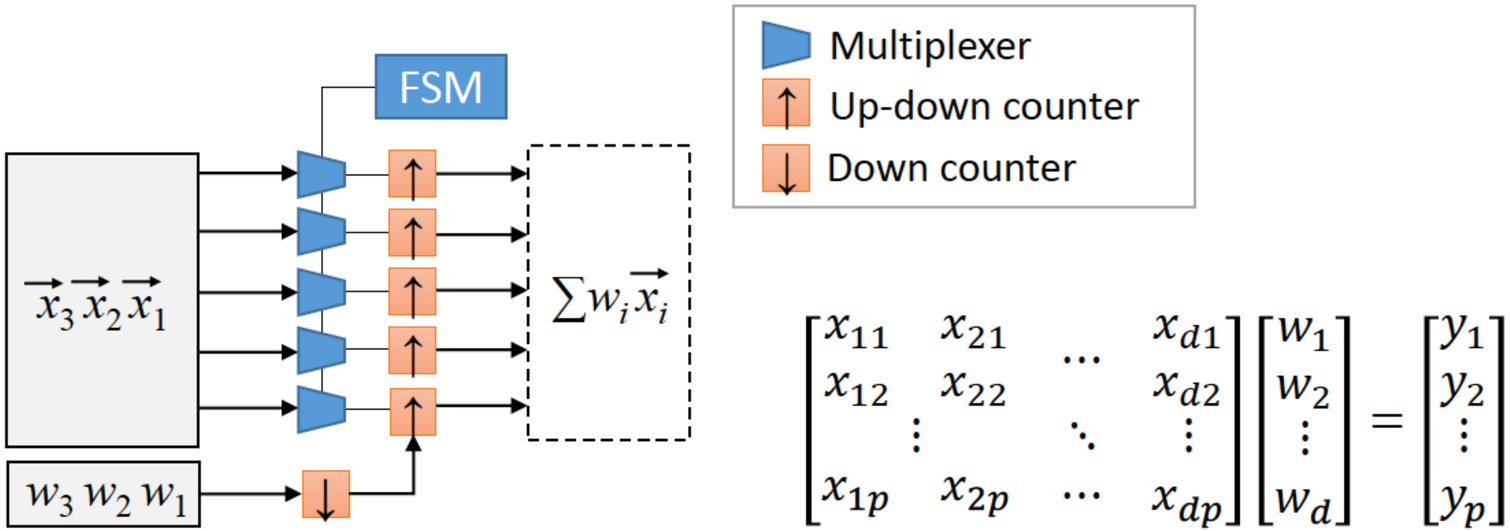
MVM Controller

- Implement fractional multiplication
- Reduce the required arithmetic unit
- Stochastic number generator



Multiplexer

- The way to implement matrix multipliers
- Synchronous operation during transmission (reduced waiting transmission time) -> Hardware acceleration



$$\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{d1} \\ x_{12} & x_{22} & & x_{d2} \\ \vdots & & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{dp} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}$$

DEMO 7

Future Work

8

"I can finally go back to take a shower."

Future work

- Parallel operation using more computing resources
- Change the model representation to reduce storage space
- Actual use of recognized voice messages (ex: robot dog)
- If you have a higher level FPGA, you can try a more complicated model.

Reference

- Chen, Yu-Hsin, et al. "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks." IEEE Journal of Solid-State Circuits 52.1 (2017): 127-138.
- Sim, Hyeonuk, and Jongeun Lee. "A new stochastic computing multiplier with application to deep convolutional neural networks." Proceedings of the 54th Annual Design Automation Conference 2017. ACM, 2017.
- Kim, Daewoo, et al. "FPGA implementation of convolutional neural network based on stochastic computing." Field Programmable Technology (ICFPT), 2017 International Conference on. IEEE, 2017.
- Alaghi, Armin, and John P. Hayes. "Survey of stochastic computing." ACM Transactions on Embedded computing systems (TECS) 12.2s (2013): 92.

Thanks!

Any questions?