fit@hcmus

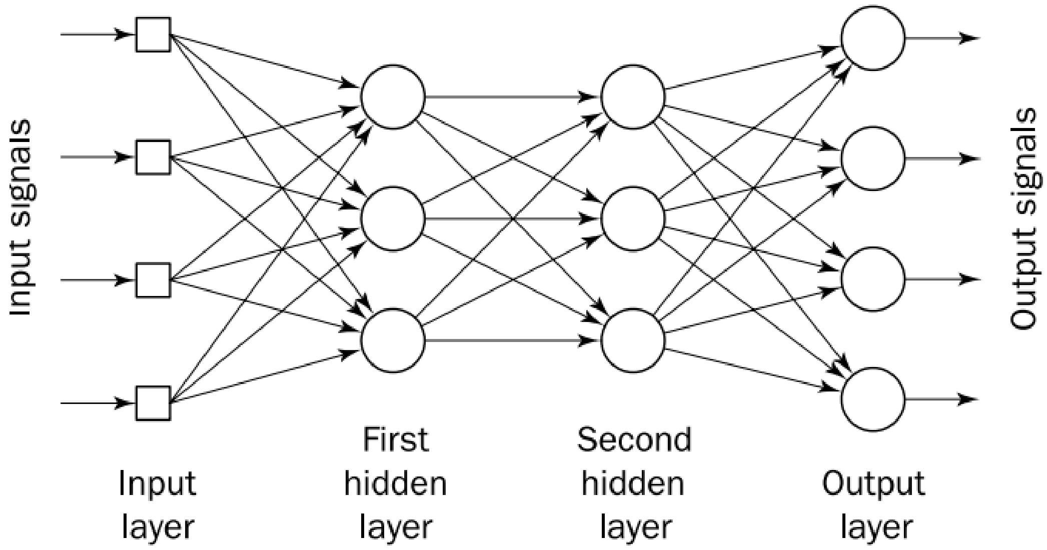
Vietnam National University, Ho Chi Minh City
University of Science
Faculty of Information Technology

CSC14120 – PARALLEL PROGRAMMING

FINAL PROJECT

1. Introduction

In this final project, you will be implementing and optimizing the traditional Artificial Neural Network (ANN) which consist of: 1 input layer, 2 hidden layers, 1 output layer.



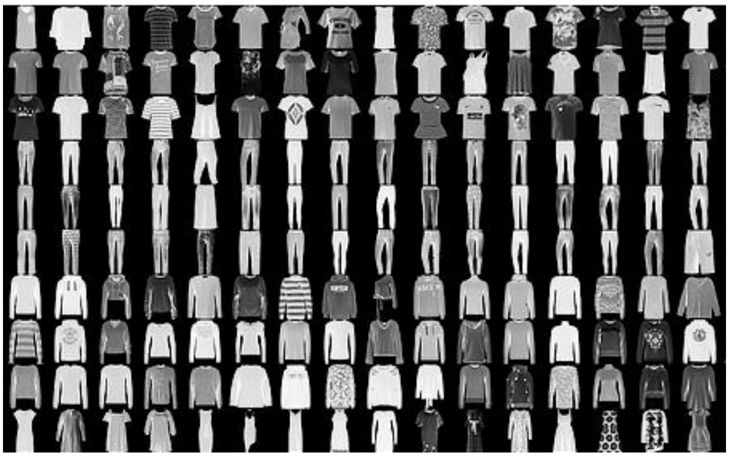
The detail architecture:

```
Layer (type)                 Output Shape              Param #
=================================================================
flatten (Flatten)            (None, 784)               0
_________________________________________________________________
dense (Dense)                (None, 128)               100480
_________________________________________________________________
dense_1 (Dense)              (None, 128)               16512
_________________________________________________________________
dense_2 (Dense)              (None, 10)                1290
```

The activation function on 2 hidden layers is ***ReLU function***

The activation function on output layer is ***softmax function***

Fashion-MNIST is a dataset of Zalando's article images—consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes. We intend Fashion-MNIST to serve as a direct drop-in replacement for the original MNIST dataset for benchmarking machine learning algorithms. It shares the same image size and structure of training and testing splits.



The overall learning objectives for this project are:

Demonstrating command of CUDA and optimization approaches by designing and implementing an optimized artificial neural network.

2. Background knowledge

- This video "What is Neural Networks" give a very good explanation with respect to ANNs. You can also check other good materials about Neural Network here from the same authors.
- This Neural Networks and Deep Learning is a free online book. It also has one chapter explain how to use neural network to recognize handwritten digits.

3. Scope

What you need to do:

Implementing and optimizing the ANN.

- Design and implement host code
- Implement a basic GPU kernel.
- Optimize your GPU kernel. Some ideas to optimize:
  - Tiled shared memory convolution
  - Shared memory matrix multiplication and input matrix unrolling
  - Kernel fusion for unrolling and matrix-multiplication
  - Weight matrix (kernel values) in constant memory
  - Tuning with restrict and loop unrolling
  - Sweeping various parameters to find best values (block sizes, amount of thread coarsening)
  - Multiple kernel implementations for different layer sizes
  - Input channel reduction: tree
  - Input channel reduction: atomics

- Fixed point (FP16) arithmetic.
- Using Streams to overlap computation with data transfer
- An advanced matrix multiplication algorithm

4. Submission

Report

As you implement your optimizations, you are required to document their effect on performance. Create a document and describe each optimization you implemented, including why you selected this optimization, show your output result, and the output of each layer's timing with this optimization. Describe in detail how you implemented the optimizations.

This should be write on *.ipynb notebook.

You will need to submit:

- Team plan and work distribution file.
- Notebook reports
- All source code file and an instruction file on how to set up and run your project.
- A presentation video about 15-20min. Upload on YouTube with Unlisted option. And give link on your report or readme file. (DO NOT Submit video on Moodle)