CDA5125 Programming Assignment No. 2: Improving Deep Neural Network Code with x86 Vector Extensions

(Due: 02/21/2022)

Purpose:

- Practice programming with x86 vector extensions.
- Experience improving a practical application (deep neural network) with x86 vector extensions.

Statement of work:

This is a group assignment. Each group can have 2 people. In this assignment, you will improve the deep neural network code that you developed in Assignment 1 by using x86 vector extensions. You can either use SSE or AVX in the assignment. You should apply the optimization to the most time consuming part of the program and thrive to achieve a noticeable improvement in overall execution time of the program.

Due dates:

The assignment is due on February 21, 11:59pm. Put all related source code, the makefile, and a README file in a tar file and submit the tar file. In the README file, you must describe (1) how to compile and run the program, (2) whether and how the improved program achieves high accuracy (optimized code should be correct), (3) which parts of the program that you apply the vector extensions, and (4) the performance improvement with the vector extensions - the speedup in the total execution time with respect to the original code. What is described in the README file must be repeatable with your submitted files.

Grading:

1. Submission has all components (all related source code, makefile, README file); the executable can be successfully produced with a 'make' command in the directory; a deep neural network for handwriting digit recognition with the MNIST dataset is built, x86 vector extensions are used in the code (30 points).
2. The README file describes the information as required (5 points).
3. Points in 3) and 4) can be obtained only after all points in 1) and 2) are obtained. One can follow the description in the README file to repeat the claims (5 points).
4. The improvement of the overall program execution time is substantial (10 points).